UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Detecção de anomalia através da comparação de modelos representativos

Giovana Jaskulski Gelatti

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Giovana Jaskulski Gelatti

Detecção de anomalia através da comparação de modelos representativos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
Maio de 2021

Giovana Jaskulski Gelatti

Anomaly detection through representative model comparison

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
May 2021

Dedicatótia,

Em especial, às pesquisadoras de computação e matemática por abrirem portas não lógicas e servirem de inspiração.

AGRADECIMENTOS

Foram anos prestando serviços e promovendo conhecimento novo a uma área que não era para mim natural quando iniciei o doutorado. Muitas situações inovadoras me foram apresentadas neste período: identificar problemas clínicos reais, ser apresentada a várias situações desafiadoras para ambas as áreas, propor soluções úteis, estar habituada aos termos médicos, apresentar técnicas computacionais aos especialistas, propor soluções aos problemas encontrados, validar com especialistas e demonstrar os resultados são alguns exemplos. Tais situações me foram apresentadas no estágio no exterior, em Portugal, onde me deparei com muitos problemas da área que propus soluções e onde encontrei o meu espaço como cientista.

Os últimos meses foram extremamente exigentes: pandemia, dados requeridos chegaram, corrida pelos resultados e atenção aos clínicos e hospitais envolvidos. Nada disso seria possível sem uma boa equipe de apoio.

Agradeço, primeiramente, ao Prof. André de Carvalho por me orientar e incentivar durante todos esses anos. Agradeço também pela humanidade no tratamento e orientação, por sempre me representar e apoiar. Junto ao meu orientador, agradeço da mesma forma ao orientador dos projetos realizados em Portugal, Prof. Dr. Pedro Pereira Rodrigues. Juntos formaram a equipe pela qual tive a honra de ser orientada. As discussões promoveram o meu desenvolvimento como pesquisadora e pessoa. Meus passos acadêmicos e profissionais são inspirados na trajetória e pessoa de vocês. Muito obrigada.

À minha família pelo suporte emocional e, muitas vezes, financeiro. Passamos um bom tempo distantes fisicamente, a distância RS-SP aumentou e tivemos um oceano de distância. As chamadas, visitas e apoio emocional me relembraram muitas vezes o motivo de iniciar o doutorado e continuar para chegar até aqui. Pai, mãe e Ju, vocês são meu porto seguro. Obrigada por me lembrarem da minha capacidade e estarem sempre "por perto"da nossa maneira. Vocês são incríveis.

À psicóloga Daniela pelos atendimentos, suporte e apoio para fortalecimento nesta importante fase.

Aos amigos de São Carlos, principalmente as mulheres com quem dividi república e colegas do ICMC. Vocês me lembraram muitas vezes que a vida deve ser aproveitada da melhor forma, com equilíbrio. Obrigada pela presença e pelos momentos divertidos.

Aos amigos do Porto, vocês são minha família em Portugal. São para sempre: nós por nós. Muitos cafés, fugidinhas para um pôr do sol e discussões acompanharam o meu desenvolvimento e "refrescaram a cabeça"sempre que precisei. Vocês viram de pertinho meu crescimento e como tudo mudou da metade para o fim deste processo. Agradeço a todos pelo suporte, onde realmente ninguém soltou a mão de ninguém. Principalmente à minha família de casa, Arthur e Olívia, que acompanharam de pertinho a emocionante finalização, essa é para vocês.

À Dr.ª Filipa Rafael pela disponibilidade de validação, discussões e auxílio na utilização dos termos clínicos e médicos e definições desta área.

À equipe VirtualCare pelo suporte e parceria em todos os momentos. Trabalhar nesta empresa me proporcionou um grande salto de descobertas da aplicação das técnicas computacionais e análise de dados clínicos. Agradeço especialmente à Eliana, pelas muitas análises e requerimento de dados realizados, e ao Tiago pela disponibilidade de tempo e recursos para o desenvolvimento desta tese. Juntamente, agradeço a Drª Priscila Maranhão pela disponibilidade e ao Ricardo Pereira pelo auxílio nas discussões e análises.

À Universidade de São Paulo por promover e disponibilizar professores de cunho científico de altíssima qualidade que me proporcionaram discussões e suporte de igual qualidade. Ao ICMC e seus serviços acadêmicos, pela estrutura e apoio mesmo distantes.

Aos projetos e agências de fomento por financiar conhecimento e apoiar pesquisadores no Brasil, ao CNPq e CAPES-PROEX pelas bolsas desta pesquisa e, em Portugal, ao Programa Operacional regional do Norte (NORTE2020) através do Acordo de Parceria PORTUGAL2020 e do Fundo Europeu de Desenvolvimento Regional (FEDER) pela possibilidade de contribuição aos projetos Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics (NanoSTIMA) e ObsCare WCR.

Enfim, agradeço a todas as mulheres que me mostraram um caminho possível, as portas não-lógicas na computação, e inspiraram este trabalho. A todas as pessoas que me colocaram no caminho e garantiram a minha permanência.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, vinculado ao projeto Desafios em IA: algoritmos, modelos e aplicações.

"I would say that the system
as a whole needs to set experiential learning
and the development of critical thinking
as a goal and engage members of
society as partners."
(Freda Yawson)

RESUMO

GELATTI, G. J. **Detecção de anomalia através da comparação de modelos representativos**. 2021. 120 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Lacuna: Regulamentos de privacidade e acesso a dados e leis de proteção de dados tornam a comparação de departamentos e identificação de padrões, em geral, tarefas difíceis. A exploração dos dados coletados, juntamente com um modelo descritivo induzido a partir desses dados, podem ajudar a identificar modelos destoantes e promover a comparação das instituições.
Objetivo: O estudo propõe a criação de modelos de redes Bayesianas capazes de extrair conhecimentos novos e significativos a partir dos dados nas variáveis utilizadas no estudo de caso. Propomos a criação de modelos de seções obstétricas por meio de variáveis utilizadas para classificação de Robson (CR), utilizada para classificar gestantes em 10 grupos, estudo de possível integração de novas variáveis à CR, a recriação de dados pelos modelos e a detecção de departamentos obstétricos com comportamentos gerais muito diferentes (anômalos) pela comparação dos modelos.
Métodos: Foi desenvolvido um modelo de rede Bayesiana com as variáveis utilizadas para CR por cada hospital envolvido no estudo. Propusemos e investigamos experimentalmente novas variáveis que podem melhor caracterizar e distribuir as gestantes nos grupos de Robson. As funções do pacote R "bnlearn" foram usadas para manipular e recriar dados no modelo. O desempenho deste modelo foi validado quanto à capacidade de recriar dados, comparando com medidas estatísticas de dados reais, e verificando se as distribuições nos grupos CR permanecem as mesmas. No intuito de construir uma matriz de distância para identificação de dados destoantes, a distância de Hamming foi utilizada para calcular as dissimilaridades entre os modelos. As anomalias detectadas foram validadas por especialista de acordo com a escala Likert.
Resultados: Os dados foram descritos e recriados através de redes Bayesianas com imputação de dados, com referência significativa aos dados reais. A comparação dos modelos sobre as seções de obstetrícia identificou padrões e anomalias. A comparação permitiu diferenciar os setores com diferentes taxas de cesárea e distribuição nos grupos de Robson, de acordo com as variáveis selecionadas, preservando o acesso aos dados reais das instituições.

Palavras-chave: Detecção de anomalia, Aprendizado não supervisionado, Modelos representativos, Redes Bayesianas.

ABSTRACT

GELATTI, G. J. **Anomaly detection through representative model comparison**. 2021. 120 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Background: Data access and privacy regulations and data protection laws make comparing departments in general and pattern detection difficult tasks. The collected data exploration together with a descriptive model induced from these data might help to identify dissonant models and promote the comparision among institutions.
Aim: This study proposes the creation of Bayesian network models which able to elicit new and meaningful knowledge from the data in the variables used in the case study. We propose to create models of obstetric departments through the variables adopted in Robson classification (RC), used to classify pregnant women in 10 groups, the study of possible integration of new variables to RC, the recreation of data by the models and the detection of obstetric departments with very different general behaviors (anomalous) by comparing the models.
Methods: A Bayesian network model was developed with the variables used for RC by each hospital involved in the study. We proposed and experimentally investigated new variables that can better characterize and distribute pregnant women in Robson's groups. The "bnlearn" R package functions were used to manipulate and recreating data across the model. The performance of this model was validated regarding its capacity of recreate data, comparing with real data statistic measures, and verifying whether the distributions on RC groups remain the same. To set up a distance matrix to identify dissonant data, Hamming distances were calculated model by model. The anomalies detected were validated by a specialist according to the Likert scale.
Results: The data were described and recreated through Bayesian networks with data imputation, with a meaningful reference to real data. By comparing obstetric departments it was possible to identify patterns and anomalies. The comparison allowed to differentiate the departments with different cesarean rates and RC groups distributions, according to selected variables, thus preserving access to the real data of the institutions.

Keywords: Anomaly detection, Unsupervised learning, Representative models, Bayesian networks.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de matriz de dissimilaridade 27
Figura 2 – Exemplo de distância entre observações num sistema de duas dimensões (x , y) 28
Figura 3 – Tipos de grafos 34
Figura 4 – Espaços amostrais A e B utilizados para demonstração das probabilidades de dois eventos 35
Figura 5 – Estrutura de uma rede Bayesiana de 2 nodos relacionando câncer metastático e cálcio sérico 38
Figura 6 – Estrutura de uma rede Bayesiana 39
Figura 7 – Estrutura de uma rede Bayesiana com probabilidades sem inferências 39
Figura 8 – Estrutura de uma rede Bayesiana com probabilidades ajustadas ao inferir probabilidade 100% ao estado *"Increased"* no nodo B 40
Figura 9 – Exemplo de anomalias em uma representação de duas dimensões. Fonte: Imagem retirada do trabalho de Chandola, Banerjee e Kumar (2009) 44
Figura 10 – Anomalias identificadas no conjunto de dados Iris com k=3 47
Figura 11 – Gráfico 3D do conjunto de dados Iris. Em vermelho, o resultado das 5 observações mais anômalas no conjunto de dados Iris identificados pela técnica LOF com k=5 49
Figura 12 – Processo de criação de modelos e amostragem de dados deles em uma instituição exemplo H1 62
Figura 13 – Rede Bayesiana produzida por meio dos dados do hospital H1 80
Figura 14 – Rede Bayesiana produzida por meio dos dados do hospital H2 80
Figura 15 – Rede Bayesiana produzida por meio dos dados do hospital H3 81
Figura 16 – Modelo gerado pelos dados do hospital H1 com dados preenchidos com média e moda das variáveis 86
Figura 17 – Modelo gerado pelos dados do hospital H2 com dados preenchidos com média e moda das variáveis 87
Figura 18 – Modelo gerado pelos dados do hospital H3 preenchidos com média e moda das variáveis 88

LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – Criação de redes Bayesianas e dados artificiais a partir de RB com funções do pacote *bnlearn* no R 56

LISTA DE TABELAS

Tabela 1 – Tabela com as probabilidades condicionais do nodo B 38
Tabela 2 – Tabela com as probabilidades condicionais do nodo D da Figura 7 39
Tabela 3 – Descrição dos grupos da classificação de Robson 60
Tabela 4 – Referência sobre a descrição dos grupos da classificação de Robson 69
Tabela 5 – Comparativo da taxa de *missings* por variável pré e após preenchimento determinístico 72
Tabela 6 – Frequência, $p-value$ e contribuição ao tipo de parto nas variáveis no total da amostra de hospitais portugueses e significância entre estas variáveis . . 74
Tabela 7 – (continuação tabela 6) 75
Tabela 8 – (continuação tabela 7) 76
Tabela 9 – Contribuição do tipo de parto nas classes de cesarianas no total da amostra de hospitais portugueses e significância entre estas variáveis 76
Tabela 10 – Contribuição do tipo de parto no número de gestações no total da amostra de hospitais portugueses e significância entre estas variáveis 77
Tabela 11 – Contribuição do tipo de parto no número de filhos no total da amostra de hospitais portugueses e significância entre estas variáveis 77
Tabela 12 – Contribuição do tipo de parto no número de cesarianas anteriores no total da amostra de hospitais portugueses e significância entre estas variáveis 77
Tabela 13 – Total de registros por hospital e total de registros completos 78
Tabela 14 – Moda das variáveis selecionadas nos conjuntos de dados reais, seleção dos registros completos e dados recriados 79
Tabela 15 – Características dos dados referentes a quantidade de bebês na gestação (variável *NUM_RN*) nos conjuntos de dados dos hospitais (média [min, max]) . 81
Tabela 16 – Tabela de contingência com distribuição da frequência das participações em porcentagem de registros nos grupos de Robson (%) no total de dados reais sem preenchimento, dos dados preenchidos pela média e moda (input1) e preenchidos com a abordagem MICE (input2) 84
Tabela 17 – Tabela de contingência com distribuição da frequência das participações em porcentagem de registros pelos grupos de Robson (%) no total de dados originais (N) e total de registros completos (n). Os dados recriados possuem o tamanho dos dados originais. 85

Tabela 18 – Distribuição da participação dos partos nos grupos de Robson dos dados que geraram o primeiro modelo considerado anômalo pela estrutura do modelo gerado conter diferentes nodos 90
Tabela 19 – Distribuição da participação dos partos nos grupos de Robson dos dados que geraram o modelo considerado anômalo pela detecção de anomalia 90

LISTA DE ABREVIATURAS E SIGLAS

AA	Aprendizagem automática
CES	Comissão de Ética pela Saúde
CR	Classificação de Robson
HC	*Hill-climbing*
KNN	Algoritmo de detecção de anomalia que utiliza o kNN
kNN	k-nearest neighbors
LGPD	Lei Geral de Proteção de Dados Pessoais
LOF	*Local Outlier Detection*
MAR	*Missing at random*
MCAR	*Missing completely at random*
MICE	*Multiple Imputation by Chained Equations*
ML	*Machine learning*
MNAR	*Missing not at random*
RAI	responsável pelo acesso à informação
RB	Redes Bayesianas
RGPD	Regulamento Geral sobre a Proteção de Dados
TABU	Tabu search
TP	Trabalho de parto
UE	União Europeia

SUMÁRIO

1 INTRODUÇÃO 25
1.1 Descrição detalhada 27
1.2 Perguntas de pesquisa e hipótese 29
1.3 Objetivo geral 30
1.3.1 Objetivos específicos 30
1.4 Estrutura e organização do trabalho 30

2 MODELOS REPRESENTATIVOS 33
2.1 Redes Bayesianas 34
2.2 Discussão 40

3 DETECÇÃO DE ANOMALIA 43
3.1 Técnicas baseadas nas distâncias 47
3.2 Técnicas baseadas na densidade 48
3.2.1 LOF 48
3.3 Medidas de distância 48
3.4 Discussão 49

4 PROPOSTA 51
4.1 Metodologia 51
4.2 Materiais e métodos 52
4.3 Anonimização dos dados 52
4.4 Falsos *missings*, mecanismos e preenchimento de dados faltantes . 53
4.5 Redes Bayesianas e estudo de caso 54
4.5.1 Validação 56
4.6 Detecção de entidades anômalas 57
4.7 Validação 57

5 CASO DE ESTUDO: DADOS CLÍNICOS 59
5.1 Contextualização 59
5.2 Metodologia 61
5.3 Base de dados 63

5.3.1 Descrição das variáveis 63
5.4 Anonimização dos dados 67
5.5 Pré-processamento dos dados 67
5.5.1 Validação dos valores e preenchimento dos falsos missings 68
5.6 Preenchimento de dados 73
5.7 Modelo representativo dos dados obstétricos 77
5.8 Detecção de anomalia 82

6 RESULTADOS 83
6.1 Preenchimento de dados 83
6.2 Modelos gerados 84
6.3 Detecção de anomalia 89

7 CONCLUSÃO E CONTRIBUIÇÕES 91
7.1 Contribuições 93
7.1.1 Publicações aceitas 94
7.1.2 Publicação aguardando revisão 94
7.2 Trabalhos futuros 94

REFERÊNCIAS 97

GLOSSÁRIO 103

APÊNDICE A QUESTIONÁRIO DE VALIDAÇÃO 105

ANEXO A AUTORIZAÇÕES DOS COMITÊS DE ÉTICA 109

CAPÍTULO

1

INTRODUÇÃO

Assim como o avanço da tecnologia, também a produção de dados cresce de maneira incontida. Analisar, medir, armazenar e fazer a coleta real desses dados pode envolver erros, incertezas e informações que podem não ser detectadas pela análise humana. A perda de conhecimento possível de ser obtido por essa razão acarreta a impossibilidade de realizar análises para que novas descobertas científicas possam gerar melhorias no funcionamento do objeto ou entidade estudada.

Por vezes, tais eventos são importantes e representam algo significativo para o estudo. Esses eventos não detectados pela análise humana podem ser identificados por técnicas de Aprendizagem automática (AA), conhecida como aprendizagem de máquina ou pelo termo em inglês *Machine learning* (ML). Nesta área, eventos raros ou fora do padrão são chamados de anomalia, e técnicas de detecção de anomalias podem identificá-los (CHANDOLA; BANERJEE; KUMAR, 2009).

A definição de anomalia em AA é, num sentido gráfico ou de distância entre observações, a "*observação que se desvia tanto das restantes que pode-se suspeitar ter sido gerada por um mecanismo distinto do mecanismo que gerou os dados de referência*"(HAWKINS, 1980). Na abordagem do projeto, é previsto que a detecção de anomalias permita identificar comportamentos e situações extremamente dissemelhantes, que estejam em discordância com o padrão que os dados possuem. Isso dará aos especialistas a capacidade de observar erros, casos raros e identificar estratégias para melhorias no comportamento de uma entidade. Uma forma de detectar as anomalias é pela comparação de entidades. Uma entidade, no contexto deste trabalho, é algo que se queira comparar, o conjunto de dados sobre um objeto de estudo. Pode ser uma instituição, dados de um indivíduo ou conjunto de dados climáticos, por exemplo.

A comparação de entidades permite que seja encontrado um padrão no comportamento

dos dados. Identificando os comportamentos, e através da comparação, é criada uma matriz de dissemelhança que quantifica a diferença entre os comportamentos das entidades, sendo utilizada nas técnicas de detecção de anomalia. Entretanto, é importante que se possua uma globalização dos dados, gerando uma maior amostra para o estudo. A importância aumenta quando, por exemplo, a entidade a ser comparada é uma doença rara. Se os dados sobre os pacientes de diferentes centros estiverem disponíveis, outros casos dessa doença poderiam ser encontrados, e também poderiam ser identificados relatos de sucesso no diagnóstico e tratamento, características do paciente entre outras informações relevantes. Desta forma, podem ser identificados grupos ou padrões de comportamentos distintos.

No entanto, por razões burocráticas e conceituais, existem barreiras a serem desconstruídas e leis a serem seguidas para que a globalização das amostras e o acesso aos dados seja realmente uma prática comum. Além disso, é necessário que seja preservada a privacidade dos indivíduos cujos dados serão compartilhados. O não compartilhamento de dados impede a análise global dos mesmos. Essa análise permitiria um avanço da ciência, por exemplo, para personalizar diagnósticos, aumentando os dados amostrais quando compartilhados entre hospitais. Por essas razões é motivado o desenvolvimento de novas estratégias para a disponibilização dos dados, respeitando a proteção do direito à privacidade.

Uma possível estratégia é a criação de modelos que representem os dados originais. Com este artifício é possível transpor uma parcela das barreiras burocráticas para aceder aos dados e preservar a anonimização destes. Sobre os modelos representativos será feita a comparação e detecção das anomalias. Ao comparar modelos, serão identificadas as entidades que não se assemelham ao comportamento padrão do todo. A identificação destas entidades, a que chamamos de anomalias, permitirá a avaliação de estratégias adotadas em determinados comportamentos que possam agregar ou causar melhoria para outras entidades.

Como caso de estudo, e para exemplificar a abordagem da investigação, foi identificado em um centro de investigação em Portugal um problema real: a impossibilidade de comparar seções de obstetrícia a fim de novas descobertas para a descrição dos serviços de obstetrícia e redução da taxa de cesariana em Portugal.

Em uma visão global, partos do tipo cesariana, além de agregarem alto custo ao procedimento, possuem um risco de vida para a grávida e o bebê (MYLONAS; FRIESE, 2015; SILVER *et al.*, 2006). Por isso, pretende-se desenvolver um modelo para cada hospital envolvido na investigação para decisão de partos (cesarianas/normais). A partir dos modelos construídos, pretende-se compará-los através de uma medida de dissimilaridade a ser integrada ao projeto e identificar os comportamentos anômalos. Os modelos podem possuir um desfecho clínico positivo (baixa taxa de partos cesarianas) ou negativo (alta taxa de partos cesarianas) e através das anomalias e seus desfechos clínicos integrados pode ser feita a avaliação do especialista nas

possíveis mudanças de comportamento que resultariam em melhorias.

Este trabalho contempla a criação de modelos representativos. A partir desses modelos é possível comparar e detectar comportamentos que diferem substancialmente dos outros. As redes Bayesianas servem para descrever o relacionamento entre as variáveis da base de dados analisada por meio de uma estrutura de nodos e arestas direcionadas e probabilidades ligadas a elas. Redes Bayesianas serão criadas localmente e os dados reais não serão alterados ou compartilhados. Esses modelos serão comparados a fim de detectar aqueles que se desviam significativamente do conjunto.

1.1 Descrição detalhada

A detecção de anomalias e/ou *outliers* é definida como detecção de observações anormais baseada em modelos predefinidos normais (WU; BANZHAF, 2010), detecção de desvio, entre outros nomes (TAN *et al.*, 2006). Para além disto, na maioria dos casos, os dados e modelos construídos não são anotados como "anômalos ou não"pelos especialistas pois isto requer alto custo de mão de obra especializada e tempo para análise. Por esta razão técnicas não supervisionadas são indicadas, resultando na utilização de técnicas de detecção de anomalias não supervisionadas para esta investigação. Um exemplo de detecção não supervisionada é a detecção de anomalia através de técnicas de agrupamento de dados. Ao agrupar os dados classifica-se o conjunto na estrutura de grupos. Cada grupo é formado pelas entidades ou observações que mais se assemelham entre si com base em uma matriz de dissimilaridade. A matriz de dissimilaridade é a quantificação de uma unidade de distância entre observações, como representado na Figura 1, onde, por exemplo, $d(2,1)$ significa a distância da observação 1 à observação 2. Transpondo em um modelo bidimensional, $d(2,1)$ representa a aresta ligando as observações 1 e 2 na representação da Figura 2. No contexto dessas distâncias, do agrupamento de dados e de forma superficial, uma anomalia pode ser definida como entidades ou observações distantes do grupo que foram associadas.

$$D = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Figura 1 – Exemplo de matriz de dissimilaridade

Ao comparar as entidades pela dissemelhança, interessa-nos descobrir as influências dos seus atributos nas possíveis decisões e a relação das entidades entre si. Num cenário de seções de obstetrícia, por exemplo, pode-se, por análise do comportamento dos serviços, identificar

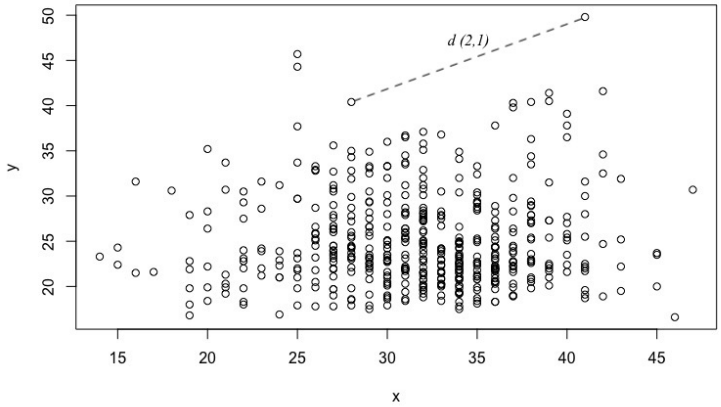


Figura 2 – Exemplo de distância entre observações num sistema de duas dimensões (x , y)

hospitais que se destacam ao ter uma taxa significativamente baixa, objetivando diminuir a taxa global de cesarianas em um hospital. Para isso, o acesso aos dados originais é necessário. No entanto, o processo de requerer, analisar, processar e compartilhar dados e resultados entre entidades não é uma tarefa simples.

Conforme o Regulamento Geral sobre a Proteção de Dados (RGPD) da União Europeia (UE) 2016/679 (REGULAMENTO. . . , 2016), os dados só podem ser compartilhados com o consentimento das entidades as quais os dados representam. Por exemplo, os dados sobre exames realizados e características de pacientes de um certo hospital devem ser autorizados por cada paciente que constitui esta base de dados, pois dele é a propriedade dos seus dados, como regia a lei sob o Artigo 3º da Lei portuguesa nº 12/2005. Apesar deste impasse, cada instituição possui um responsável pelo acesso à informação (RAI) que permite o acesso aos dados, interpreta as leis, a doutrina dos órgãos superiores de proteção aos dados e jurisprudência do Supremo Tribunal Administrativo (Artigo 89 nº2 de 2016 do Regulamento Geral sobre a Proteção de Dados da UE). O RAI pode conceder dados por decisão própria regido pela lei de acesso e utilização 26/2016 que teve seu início de aplicabilidade em maio de 2018. Nesta lei, dados de saúde podem ser compartilhados sem o consentimento dos titulares dos dados para fins de investigação científica, desde que anonimizados e com autorização do responsável pelo acesso à informação. Este processo ocorre em acordo conjuntamente com a autorização pelo comitê de ética de cada instituição, seguindo Lei Geral de Proteção de Dados Pessoais (Lei n.º 58/2019, 151/2019). Este é um processo obrigatório para o pedido de acesso aos dados e a cada novo pedido, ou seja, mesmo que os dados e a proposta forem de igual abordagem, o pedido deve ser refeito.

No Brasil, similar ao Regulamento da UE, a Lei Geral de Proteção de Dados Pessoais

(LGPD), (UNIÃO SEÇÃO 1, 2018) Lei nº 13.709, de 14 de agosto de 2018, permite o tratamento de dados pessoais "para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais".

Para além da barreira jurídica e burocrática, existem várias vertentes de investigação sobre outras barreiras e razões para que a partilha de dados não seja feita. Muitos estudos apontam a insegurança de investigadores na avaliação dos seus resultados e o uso de resultados sobre bases de dados em colheita ou ainda em desenvolvimento (WICHERTS; BAKKER; MOLENAAR, 2011; WALPORT; BREST, 2011; HUTSON, 2018). Existe também a suposição da preocupação de governos e agências para que seus dados não sejam expostos ou manipulados (PISANI; ABOUZAHR, 2010). Para além destas razões, ainda há dados que são dispensados ignorando o seu valor agregado. Essas instituições acreditam numa falsa segurança ao não compartilhar suas informações. Entretanto, enquanto os dados são localmente armazenados e analisados ou simplesmente dispensados, existe a perda de informação e de conhecimento que poderiam ser obtidos a partir destes dados. A perda deve-se ao bloqueio do acesso e partilha entre entidades e comunidade acadêmica para diferentes análises e visões que produziriam conclusões distintas. Estas prováveis descobertas são infinitas e com elas centros e instituições poderiam comunicar, relatar diferenças entre comportamentos, crescer economicamente, entre muitos outros benefícios.

As discussões sobre barreiras burocráticas, necessidade de boas práticas e desafios existentes continuam recorrentes e, enquanto estes desafios estão longe de serem resolvidos, outras soluções podem ser adotadas. Uma solução possível e viável é recorrer a técnicas computacionais de aprendizagem automática. Através de técnicas de AA pode-se recriar os dados localmente a partir dos originais sem infringir leis ou gerar burocracia em cima disso. Isto é possível e pode ser realizado, por exemplo, através da construção de modelos de redes Bayesianas(FRIEDMAN; GEIGER; GOLDSZMIDT, 1997). Ao construir os modelos de comportamento dos dados pode-se recriar dados que representem as características dos originais, comparar o comportamento de entidades e até mesmo identificar melhorias pela comparação. Foram encontrados estudos sobre as detecções de anomalias e sobre os modelos, porém na literatura não foi encontrada em revisão a identificação das anomalias na comparação entre modelos (redes Bayesianas). Essas questões são a base para o desenvolvimento do projeto e serão detalhados os estados da arte de cada contexto nas seções correspondentes.

1.2 Perguntas de pesquisa e hipótese

As perguntas que esta tese pretende responder são:

- Dados criados a partir de modelos representativos de dados reais capacitam a comparação entre diferentes bases de dados sem comprometer a fidelidade aos dados reais?
- É possível a comparação de entidades e detecção de anomalias de forma não supervisionada, sem a partilha de dados entre entidades?

E, para responder a estas perguntas, temos as hipóteses de que:

- Os dados criados através de um modelo de Bayes representam os dados originais.
- A comparação entre modelos representativos de diferentes fontes permite a identificação daqueles significativamente destoantes aos demais com relevância à área específica dos dados.

1.3 Objetivo geral

O objetivo geral do projeto é detectar entidades anômalas pela comparação sem compartilhar dados entre elas. Para que este objetivo seja alcançado, o mesmo é pormenorizado em objetivos específicos, definidos a seguir.

1.3.1 Objetivos específicos

- Recriar informações e comportamentos de uma entidade por meio de rede Bayesiana;
- Permitir a comparação sem troca de informação sobre os dados reais entre instituições/entidades;
- Proteger a análise distribuída de detecção de anomalia;
- Identificar anomalias com relevância para especialistas nos dados e para o contexto da aplicação.

A seguir são detalhados os referenciais teóricos que capacitam a resposta das questões da pesquisa e serviram como base para as escolhas de métodos utilizados.

1.4 Estrutura e organização do trabalho

Para realizar a revisão bibliográfica, foram escolhidos como mecanismos de busca as ferramentas Pubmed, Scopus e ISI Web of Knowledge. A escolha pelo Pubmed deve-se a sua

grande utilização como ferramenta de busca por publicações da área biomédica e por ser o meio de busca referência na área da saúde, área em que é realizado o caso de estudo inicial. O Scopus reúne artigos gerais e é muito utilizado no meio acadêmico. E o ISI Web of Knowledge disponibiliza uma ampla busca por artigos, capítulos de livros e citações também muito utilizados na comunidade acadêmica.

As palavras-chave de cada capítulo de fundamentação teórica são procuradas no título, no resumo e nas palavras-chave dos estudos. Para selecionar as pesquisas identificadas, foram considerados artigos científicos, publicados em revistas ou em congressos, teses, dissertações, livros e relatórios relacionados com o problema da não partilha de dados, modelos representativos, detecção de anomalia e medidas de comparação/semelhança entre redes Bayesianas, conforme foi requerido por cada capítulo e etapa de desenvolvimento. Foram pesquisados artigos nos idiomas inglês e português e adicionadas à pesquisa as indicações de artigos, livros e revisões técnicas feitas por especialistas da área.

O Capítulo 2 refere-se à busca sobre os modelos representativos de dados. Nele são identificados trabalhos relevantes que utilizam modelos representativos e a definição desses tipos de modelos.

As técnicas de detecção de anomalias são definidas e detalhadas no Capítulo 3, com a descrição das técnicas e algoritmos e a comparação entre os trabalhos.

Ao agregar as técnicas de detecção de anomalias e modelos, a proposta é definida no Capítulo 4 com os trabalhos relacionados. A proposta é aplicada em um caso de estudo específico no Capítulo 5 e os resultados de cada etapa dos casos de estudo são apresentados no Capítulo 6. Por fim, as conclusões gerais sobre o trabalho e contribuições são apresentadas no Capítulo 7.

CAPÍTULO

2

MODELOS REPRESENTATIVOS

"Leaving aside the terms "law" and "theory," I distinguish principles, specific conditions, models, hypotheses, and generalizations. I argue that scientists use designated similarities between models and aspects of the world to form both hypotheses and generalizations." (Giere (2004))

Um modelo é o resultado do processamento e aprendizagem de dados reais por meio de técnicas de ML (*machine learning*) que, de alguma forma, caracterizam os dados com regras, números ou estruturas. Técnicas de ML, como agrupamento de dados e grafos - e como exemplo específico de um tipo de grafo, as Redes Bayesianas (RB) - são muito utilizadas por representarem os dados e agregarem uma estrutura à análise. Os grupos são os modelos representativos resultantes de técnicas de agrupamento de dados e os grafos são modelos representativos em estrutura de nodos com probabilidades associadas que resultam, por exemplo, numa rede Bayesiana.

As RB são adotadas frequentemente para resolver problemas de previsão ou para entendimento da relação entre os dados (PEARL, 1988; LAURITZEN, 1996). São numerosamente empregadas em estudos da área médica, como na classificação e previsão de tumor e câncer em mulheres (ANTAL *et al.*, 2003; AYER *et al.*, 2010), na previsão de riscos em cesarianas (LOVELL *et al.*, 1997; SMITH *et al.*, 2004; HOUSSEINI *et al.*, 2009) e na previsão de trabalho de parto normal após a primeira cesariana (METZ *et al.*, 2013; HARPER *et al.*, 2013; MARDY *et al.*, 2016). Por associar estrutura e probabilidades e poderem inferir conhecimento, como será visto a seguir, as redes Bayesianas são amplamente estudadas e utilizadas.

2.1 Redes Bayesianas

As redes Bayesianas são um tipo específico de grafo. Grafo é uma estrutura utilizada para modelar a relação entre objetos. Um grafo é composto por nodos e arestas e é definido como:

$$G = \{V, A\},$$

onde,

- $V = \{v_1, v_2, ...v_n\}$, define um conjunto de nodos;
- $A \subseteq V(G) \times V(G)$, define um conjunto de arcos, também chamados de ligações ou arestas, que ligam nodos.

As ligações entre os nodos são pares não ordenados que identificam a relação entre dois nodos. As arestas podem ser direcionadas ou não, o que tipifica um grafo não direcionado (Figura 3a) ou um grafo direcionado (Figura 3b).

(a) Grafo cíclico não direcionado (b) Grafo acíclico direcionado

Figura 3 – Tipos de grafos

Dependendo da relação entre os nodos e arestas, o grafo pode ser cíclico ou acíclico. A Figura 3 também demonstra o grafo cíclico não direcionado onde a relação entre os nodos é fechada e as arestas são não direcionadas (Figura 3a). A Figura 3b representa um grafo acíclico direcionado pois as arestas não formam ciclo entre os nodos e são direcionadas.

Redes Bayesianas são "estruturas gráficas para representar as relações probabilísticas entre um grande número de variáveis e fazer inferência probabilística com essas variáveis" (NEAPOLITAN *et al.*, 2004). É um método que agrega teoria de grafos, propondo uma estrutura e identificando relações causais, com a teoria das probabilidades. Na ciência da computação, o desenvolvimento de redes Bayesianas foi impulsionado por pesquisas em inteligência artificial, que visava produzir uma estrutura prática para o raciocínio de senso comum por meio do Teorema de Bayes.

O Teorema de Bayes, também chamado anteriormente de probabilidade de causas ou probabilidade inversa, é a probabilidade de um evento ocorrer baseado em um conhecimento relacionado a ele (MURRAY, 1975) , ou seja, uma probabilidade que condiciona dois eventos.

Para exemplificar, define-se os eventos A e B, representados na Figura 4, pelos espaços amostrais A e B. A probabilidade dos dois eventos acontecerem é a probabilidade da intersecção dos dois espaços, ilustrado pelo espaço com listras da Figura 4. Sabendo que a probabilidade dos dois eventos acontecerem ($P\{A \cap B\}$) é a multiplicação das probabilidades de A e B (Equação 2.1), a $P\{A \cap B\}$ é equivalente a $P\{B \cap A\}$. Além disso, se os eventos A e B são dependentes, ou seja, o acontecimento do evento A fornece informações sobre o evento B ocorrer, diz-se então que "o evento A está condicionado ao evento B". Esta probabilidade é a probabilidade condicional entre A e B, a probabilidade de A acontecer dado que B acontece, e é definida na Equação 2.2.

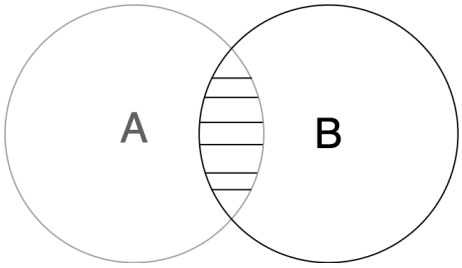


Figura 4 – Espaços amostrais A e B utilizados para demonstração das probabilidades de dois eventos

$$\mathrm{P}\{A \cap B\} = \mathrm{P}\{A\}\mathrm{P}\{B\} \tag{2.1}$$

$$\mathrm{P}\{A|B\} = \frac{\mathrm{P}\{A \cap B\}}{\mathrm{P}\{B\}} \tag{2.2}$$

Sabendo que $P\{A \cap B\}$ é igual a $P\{B \cap A\}$. Ao isolar estes termos nas equações de P{A|B} e P{B|A} temos que:

$$\mathrm{P}\{A|B\} = \frac{\mathrm{P}\{A \cap B\}}{\mathrm{P}\{B\}}$$

$$\mathrm{P}\{A|B\}\mathrm{P}\{B\} = \mathrm{P}\{A \cap B\}$$

e

$$\mathrm{P}\{B|A\} = \frac{\mathrm{P}\{A \cap B\}}{\mathrm{P}\{A\}}$$

$$\mathrm{P}\{B|A\}\mathrm{P}\{A\} = \mathrm{P}\{B \cap A\}$$

Ao igualamos os termos, temos a interpretação do Teorema de Bayes representado pela Equação 2.3.

$$\mathrm{P}\{A|B\} = \frac{\mathrm{P}\{A\}\mathrm{P}\{A \cap B\}}{\mathrm{P}\{B\}} \tag{2.3}$$

onde

- $\mathrm{P}\{A|B\}$ probabilidade de A ocorrer, dado que B acontece,
- $\mathrm{P}\{A\}$ probabilidade de A ocorrer, dado que A é um espaço amostral,
- $\mathrm{P}\{B\}$ probabilidade de B ocorrer, dado que B é um espaço amostral,
- $\mathrm{P}\{A \cap B\}$ probabilidade de A e B ocorrerem.

Interpreta-se o Teorema de Bayes a partir da Equação 2.3 , onde se sabe que B ocorre, sendo B o novo espaço amostral para o experimento aleatório onde A é um evento possível. Como na Figura 4, onde $\mathrm{P}\{A \cap B\}$ é a probabilidade de A no espaço conjunto entre A e B (espaço listrado), e a probabilidade de B é o espaço inteiro de B.

A utilização deste teorema chega a resultados com a estatística baseada no processo de aprender com as evidências. Ao inferir evidência de relações entre variáveis e probabilidades condicionais, podemos produzir uma rede Bayesiana. Desta forma, uma rede é um grafo acíclico de probabilidades condicionais. Ela é composta por dois tipos de componentes: nodos e arcos. Cada nodo representa uma variável. Os arcos ligam nodos identificando os nodos relacionados, ou seja, que possuem uma ligação condicional entre eles. Cada nodo representa uma variável com estados mutuamente exclusivos e possui probabilidades condicionais dos estados ocorrerem ou não dado a relação entre os nodos. Se o nodo não possui arcos direcionados, a probabilidade que se encontra no nodo é a probabilidade a priori. Já nos nodos com arcos direcionados a si, encontram-se as probabilidades do nodo dado que se sabe o(s) nodo(s) que possui(em) ligações com o mesmo.

Podemos exemplificar estas questões mapeando a relação entre um sinal e uma doença pelo Teorema de Bayes. Queremos saber se, dado que um paciente possui um sinal B, por exemplo, cálcio sérico elevado, a probabilidade de obter uma doença A, como um câncer metastático. Esta questão é calculada pela equação 2.4:

$$\mathrm{P}\{cancer|calcio_serico\} = \frac{\mathrm{P}\{cancer\}\mathrm{P}\{cancer \cap calcio_serico\}}{\mathrm{P}\{calcio_serico\}},$$

onde

- P$\{cancer\}$ probabilidade de uma pessoa ter câncer,
- P$\{calcio_serico\}$ probabilidade de cálcio sérico elevado,
- P$\{cancer \cap calcio_serico\}$ probabilidade de câncer e cálcio sérico,
- P$\{cancer|calcio_serico\}$ probabilidade de câncer ocorrer, dado que há cálcio sérico.

Com o teorema, podemos então pensar na rede Bayesiana. Define-se que, uma rede Bayesiana (RB) como

$$\text{RB} = (\text{G,P}), \tag{2.4}$$

onde ,

$$\text{G} = (\text{V(G)}, \text{A(G)})$$

é um grafo acíclico, com

- V(G) $= \{v_1, v_2, ...v_n\}$, define um conjunto de nodos,
- A(G) $\subseteq$ V(G) $\times$ V(G), define um conjunto de arcos ou ligações

e

- P : V(G) $\rightarrow [0,1]$ é uma distribuição de probabilidade conjunta:
  P(V(G)) : $\prod_{i=1}^{n} \text{P}(v_i \mid \pi_G(v_i))$
  $onde \pi_G$ (v_i) indica o conjunto de ligações/relações imediatas do nodo v_i em G.

Transformando o problema anterior para a estrutura da rede de Bayes, temos o conjunto de nodos V(G) = $\{v_A, v_B\}$ e mapeamos a relação com o arco direcionado do nodo A para o nodo B, sobre a variável cálcio sérico, representando a causalidade entre doença e o resultado de um exame. A estrutura desta rede é demonstrada na Figura 5.

Os nodos possuem as probabilidades de cada estado do nodo acontecer. Como exemplo, em A tem-se "*present*"e "*absent*"(presente/ausente) e em B os estados "*increased*"e "*not increased*"(elevado/não elevado). Temos que a probabilidade de uma pessoa ter câncer, a priori, é de 20%. Logo a probabilidade de não ter é de 80%. Estas são as probabilidades que se encontram no nodo A, que não possui arco direcionado a ele. No nodo B, que possui um arco direcionado a ele representando uma relação condicional (Tabela 1), encontram-se os estados de B (*increased/not increased*) condicionados ao estados do nodo A (*present/absent*). As probabilidades encontradas

no nodo B, como exemplo, são as probabilidades de B dado que A aconteceu ou que se sabe de A.

Figura 5 – Estrutura de uma rede Bayesiana de 2 nodos relacionando câncer metastático e cálcio sérico

Tabela 1 – Tabela com as probabilidades condicionais do nodo B

A	Present	Absent
Increased	0,8	0,2
Not increased	0,2	0,8

Ao acrescentar outras evidências à rede inicial da Figura 5, temos a rede representada na Figura 6. As evidências e probabilidades podem ter vindo de diferentes fontes ou aprendidas com outros dados que não os utilizados para gerar a rede anterior. Além disso, as dependências, na forma dos arcos direcionados, também podem ser inferidas por um especialista nos dados por evidência observacional.

Nesta nova rede da Figura 6, além dos nodos A e B, são acrescentados os nodos C, D e E que recebem arcos direcionados a si conforme mostra a figura. O nodo A não recebe arcos e o E não direciona arcos. Desta forma, no nodo A as probabilidades a priori continuam as mesmas. Os outros nodos ganham as probabilidades condicionadas aos nodos que cada um recebe do arco direcionado. Ou seja, C está condicionado a A, D está condicionado a B e C, e E está condicionado a C.

Na Figura 7 mostram-se as probabilidades a priori de A e as condicionais de cada nodo. Os estados em verde são os estados com maior probabilidade, como no nodo B a maior probabilidade é "*not increased*"de 68%. O nodo B continua condicionado apenas ao A, por isso mantém-se as probabilidades como na Tabela 1. Já o nodo D, recebe a o arco direcionado de B e C e tem as probabilidades na Tabela 2 com seus estados condicionados aos estados de B e C.

Diz-se que as probabilidades dissipam-se aos ramos que possuam arcos direcionados. Por isso, se inferirmos a certeza de que, em algum caso, foi verificada a certeza de um estado, por exemplo de "*increased*"do nodo B, as probabilidades dos nodos que possuem ligação para

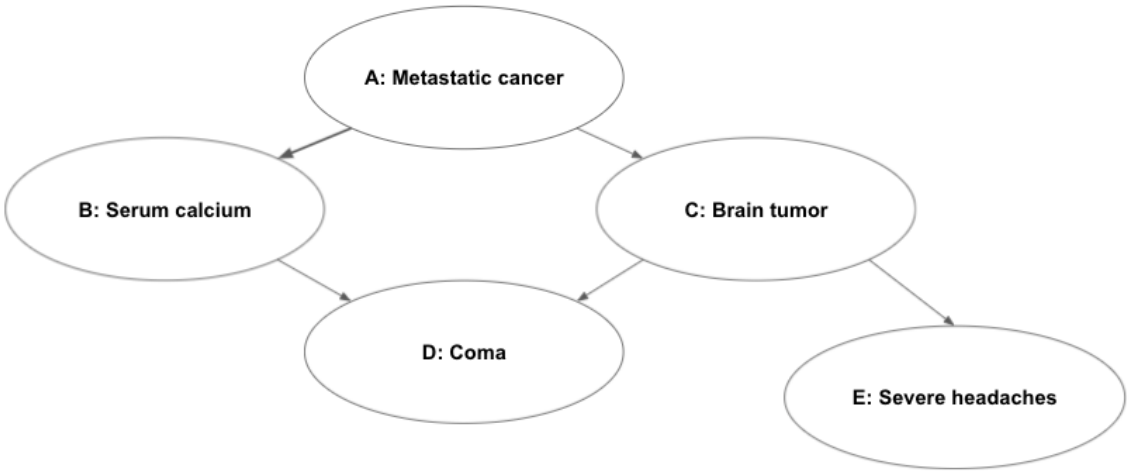


Figura 6 – Estrutura de uma rede Bayesiana

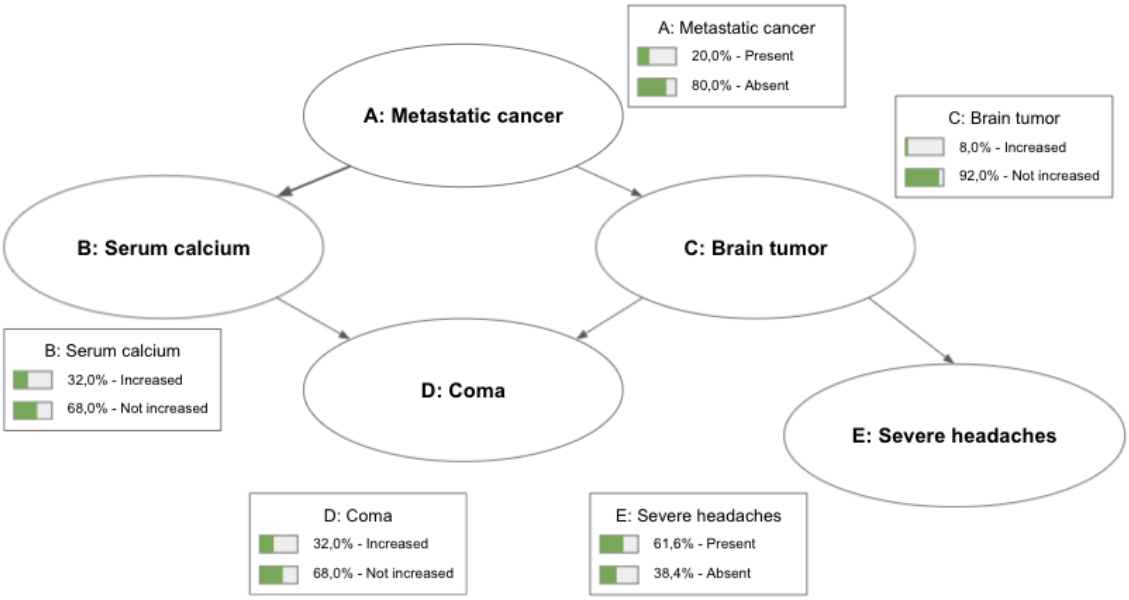


Figura 7 – Estrutura de uma rede Bayesiana com probabilidades sem inferências

Tabela 2 – Tabela com as probabilidades condicionais do nodo D da Figura 7

C	Present		Absent	
B	Increased	Not increased	Increased	Not increased
Present	0,8	0,8	0,8	0,05
Absent	0,2	0,2	0,2	0,95

ele ou dele serão ajustadas, como demonstra a Figura 8. Desta forma, a inferência é de simples manipulação e a identificação da mudança dos outros estados é verificada facilmente. Essa simplicidade auxilia o apoio a decisão em diferentes áreas.

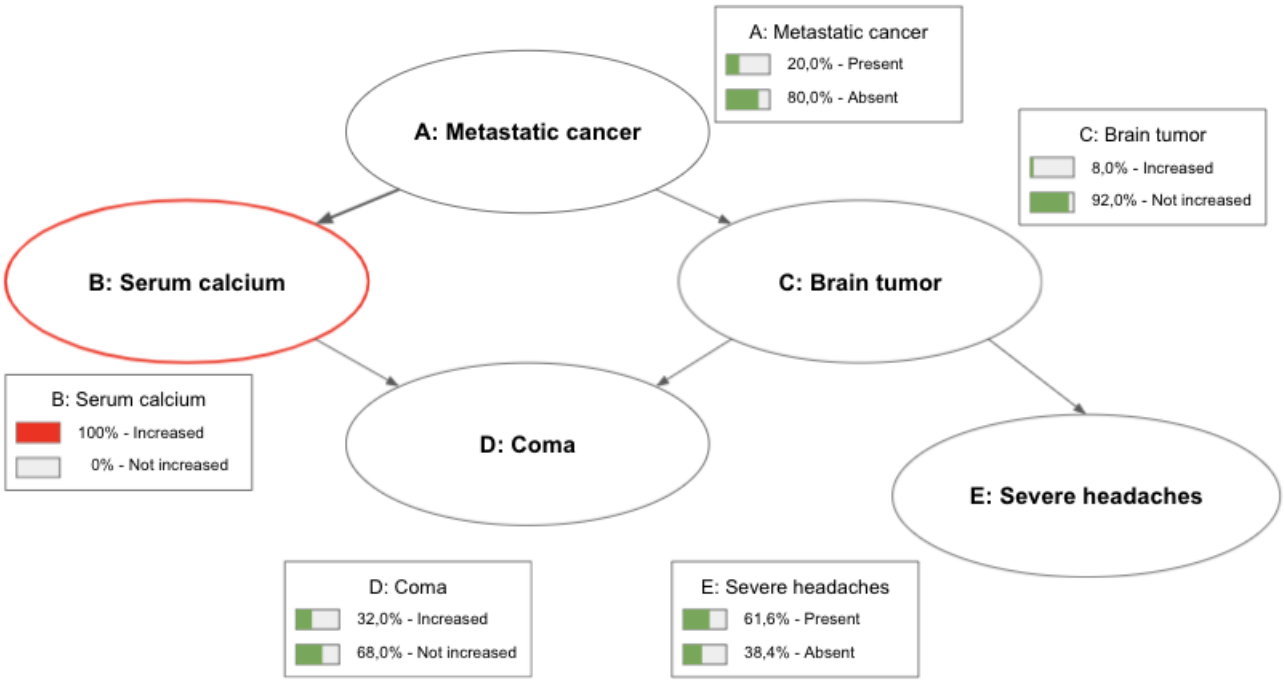


Figura 8 – Estrutura de uma rede Bayesiana com probabilidades ajustadas ao inferir probabilidade 100% ao estado *"Increased"* no nodo B

Ao inferir conhecimento a partir da rede Bayesiana, com novos nodos e probabilidades associadas, inicia-se a inferência Bayesiana com uma distribuição anterior para os parâmetros do modelo (NEAL, 1996). Com isso, não são perdidas as evidências anteriores sobre o relacionamento dos nodos que está sendo modelado.

2.2 Discussão

Nos trabalhos atuais, as redes Bayesianas foram estabelecidas como uma ferramenta onipresente para modelagem e raciocínio sob incerteza. Além disto, permite que não seja preciso reconstruir a análise ou desenvolver novos algoritmos para a análise para inferir novo conhecimento. A inferência em redes Bayesianas permite inserir novos nodos ou alterar as probabilidades dos nodos sem ter de criar novos algoritmos especializados para cada nova aplicação (DARWICHE, 2010).

Também permite a não necessidade do acesso aos dados reais para retirar conhecimento novo. É possível recriar bases de dados a partir dos modelos e também extrair conhecimento da própria rede sobre as relações entre as variáveis.

Principais vantagens:

- Transparência na análise por meio das probabilidades relacionadas;

- Permite gerar dados a partir do modelo;
- Proteção da análise distribuída;
- Análise estrutural facilitada na relação entre as variáveis e na forma de descendência entre elas;
- Possibilidade de incluir conhecimento de outras fontes, como evidência médica sobre um evento diferente observado ou probabilidades de ocorrência.

Em resumo, uma rede Bayesiana pode ser criada a partir dos dados ou a partir de probabilidades observadas, e a relação entre variáveis identificada pelo especialista. Ainda há a possibilidade de modificar a rede gerada, adicionando novos nodos ou alterando a relação e até as probabilidades, proporcionando ao especialista dos dados modificar conforme a realidade observada ou conforme novas pesquisas sugerirem e o especialista quiser acatar as estas alterações.

CAPÍTULO

3

DETECÇÃO DE ANOMALIA

"If we do not share our stories and shine a light on inequities, things will not change."
(Ellen Pao)

A detecção de anomalia tem por objetivo identificar observações que são diferentes do restante dos dados. Uma anomalia pode ser definida como uma observação que se desvia de tal forma das demais observações que se suspeita ter sido gerada por um mecanismo distinto a elas (HAWKINS, 1980). Diferentes domínios, como aprendizado de máquina, mineração de dados e estatística, utilizam termos distintos para denotar anomalia (CHANDOLA; BANERJEE; KUMAR, 2009; AGGARWAL, 2013). Alguns desses termos são: anormalidade, dado discordante, exceção aos dados, dado desviante e *outlier*. Anomalia e *outlier* são os mais frequentes.

As causas mais comuns de sua ocorrência são dados de classes diferentes rotulados com a mesma classe, valores extremos para um atributo preditivo que destoam de uma variação natural e coleta ou medição incorreta de dados (TAN *et al.*, 2006). Podem significar fraude nas telecomunicações e cartões de crédito/débito, intrusão na rede (LI, 2010; YIN; KANG; KIM, 2016; RAYANA; AKOGLU, 2016), falhas em sistemas críticos de segurança e vigilância militar para identificação de atividades de forças inimigas, num contexto clínico *"pode sinalizar um paciente em particular, sintomas não comuns ou resultados de testes que indicam uma potencial complicação em relação a saúde do paciente"* (CHANDOLA; BANERJEE; KUMAR, 2009; FACELI *et al.*, 2011) ou melhora surpreendente, e-mails que são *spams*, distúrbio em ecossistemas e presença de células raras no sangue (TAN *et al.*, 2006; AGGARWAL, 2013), entre outros tantos exemplos. Sendo assim, a identificação desse tipo de dado pode representar o ponto chave de uma solução, como um novo fenômeno para ser investigado (TAN *et al.*, 2006) .

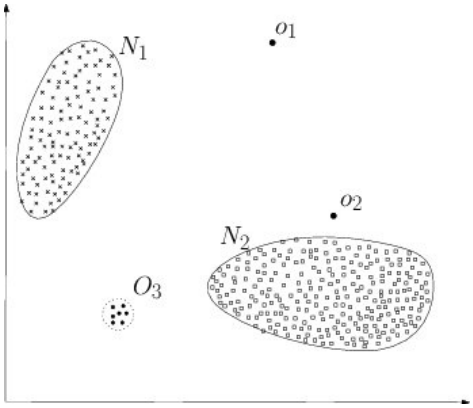


Figura 9 – Exemplo de anomalias em uma representação de duas dimensões. Fonte: Imagem retirada do trabalho de Chandola, Banerjee e Kumar (2009)

Inicialmente esses dados eram considerados como ruídos e a detecção deles servia apenas para a limpeza dos dados em uma etapa de pré-processamento. O conceito de anomalia pode ser confundido com o de ruído e novidade. Apesar de relacionados, são conceitos distintos que precisam ser claramente definidos para evitar equívocos.

Um ruído é um dado que foi registrado erroneamente, como um equívoco na coleta dos dados, um erro aleatório ou variação em um atributo preditivo medido. Ruídos devem ser removidos antes da detecção de *outliers*. A detecção de ruídos permite sua remoção, o que melhora a qualidade dos dados e da análise realizada (TAN *et al.*, 2006). Ao contrário da detecção de ruídos, a detecção de anomalias considera que dados destoantes podem representar informações relevantes que não se assemelha ao esperado ou normal observado e, por isso, ao invés de removidos, devem ser destacados. A detecção de novidades pode ser vista como um passo seguinte à detecção de anomalias. Quando uma grande quantidade de anomalias semelhantes são detectadas ao longo do tempo, elas podem ser uma novidade, devendo assim ser consideradas um novo padrão (CHANDOLA; BANERJEE; KUMAR, 2009). A incorporação de novos padrões à descrição normal dos dados é a principal diferença entre detecção de novidades e detecção de anomalias. Por isso, as técnicas utilizadas para lidar com essas tarefas são muitas vezes semelhantes àquelas utilizadas para detecção de *outliers*. A relação com os padrões identificados determinam o tipo de anomalia como anomalias pontuais, contextuais ou coletivas (CHANDOLA; BANERJEE; KUMAR, 2009).

As anomalias pontuais são detectadas na relação da observação com o restante do conjunto de dados analisado e podem ser locais ou globais. Essas observações estão ilustradas na Figura 9. A observação $O1$ se distancia de todo o restante das observações e por isso representa uma anomalia pontual global. Já a observação $O2$ poderia estar classificada no padrão N2, uma vez que é relativamente similar (menor distância). No entanto, na visão de $N2$ ela não é compatível com seu padrão normal, possuindo uma dissimilaridade maior que a suportada e por

isso é classificada como anomalia pontual local.

Se há necessidade de exemplificar o cenário no qual a observação deve ser classificada como anomalia, objetiva-se encontrar uma anomalia contextual. Ela ocorre quando determinado fenômeno está presente. O contexto normalmente é induzido pela estrutura do conjunto de dados e precisa ser especificado como parte do problema. As observações podem representar uma anomalia em um contexto e uma observação normal em outro.

Outra possível ocorrência é as anomalias serem semelhantes entre si. Neste caso, é formado um grupo ou coleção de anomalias. Ela está ilustrada na Figura 2 como o grupo O3 que não se parece com o padrão dos grupos N1 e N2 (muitas observações, de forma elipsoidal). Isso acontece, por exemplo, quando um remetente de e-mail considerado spam continua a enviar e-mails continuadamente. Os grupos anômalos também podem ocorrer pela característica dos dados quando existe correlação entre atributos (CHANDOLA; BANERJEE; KUMAR, 2009). Para detectar grupos de anomalias, deve-se atentar ao comportamento dos grupos de objetos, não somente em observações únicas.

As técnicas de detecção de anomalias podem ser desenvolvidas com base em técnicas de identificação de valores extremos, modelo de agrupamento, modelo baseado em distância, modelo baseado em densidade, modelo probabilístico e modelo de teoria da informação (TAN *et al.*, 2006; CHANDOLA; BANERJEE; KUMAR, 2009; AGGARWAL, 2013). Essas técnicas são utilizadas em um domínio abrangente de aplicação e em diferentes tipos de dados. Existem técnicas que abrangem dados de fluxo contínuo, como as aplicadas nos trabalhos de intrusão de rede; dados temporais (SUN *et al.*, 2015; BRUNO; GARZA, 2010), como de medição do clima; dados individuais de informação factual usados para fins de análise, como dados representativos de uma instituição; dados coletados sobre pacientes; medições variadas de um único evento, como os conjuntos de dados que esta tese abrange.

As técnicas aplicadas nos conjuntos de dados nos quais este trabalho se baseia identificam anomalias em apenas uma execução, não permitindo um mascaramento de anomalias quando existe uma grande quantidade de *outliers*, escondendo a presença de sua totalidade. No entanto, corre-se o risco de observações anômalas serem classificadas como normais, chamadas de *Swamping* (HADI, 1992; TAN *et al.*, 2006).

As técnicas de detecção de anomalia abrangidas aqui objetivam encontrar as exceções nos dados que podem gerar conhecimento novo e significativo como complemento de uma análise. Por isso, mesmo com o risco de *Swamping*, assume-se que as anomalias são uma pequena parte das observações e que todos casos identificados são casos raros ou exceções que merecem atenção e análise aprofundada.

As exceções podem representar, por exemplo, pacientes com características de indivíduos

saudáveis, mas que possuem como diagnóstico doente (ou o inverso) ou pacientes que possuem comportamento muito diferente de todos os outros, representando uma nova situação identificada e que merece uma análise e acompanhamento mais específicos. Tal identificação em um grande volume de dados pode ser considerada uma tarefa difícil pelo tempo e mão de obra especializada requeridos.

As técnicas de ML (machine learning) para detectar estes eventos, observações ou entidades anômalas nos conjuntos de dados podem se basear ou não no rótulo ou anotação das observações se são anomalias ou dados normais. As técnicas que utilizam um rótulo são chamadas de técnicas supervisionadas. Porém, a utilização de rótulos dos dados que identificam as observações como anomalia não representa um cenário real de geração de dados. As técnicas que não utilizam esses rótulos são chamadas de técnicas não supervisionadas.

Detectar anomalias com técnicas não supervisionadas significa detectar objetos ou eventos raros, previamente não identificáveis, sem qualquer conhecimento prévio sobre os dados ou o que significa a anomalia para o contexto. Sabe-se que, como são eventos raros, fazem parte de uma pequena parcela do conjunto de dados. Esta quantidade pode ser considerada 1% dos dados.

Definição: Dado um conjunto de dados D, a menor parcela de observações que são diferentes ou desviam significativamente ao padrão observado em D são anomalias.

Então, para realizar a detecção, é necessário quantificar a (dis)similaridade entre as observações ou entidades comparadas. Após escolhida a métrica de similaridade (distância entre entidades, por exemplo) é construída a matriz de dissimilaridade onde encontram-se todas as medidas de similaridade entre todas as entidades. Por meio desta matriz é possível agregar uma análise visual dos padrões e anomalias. Na análise computacional, o resultado das detecções de anomalias podem ser pontuações que quantificam o quão anômala é a observação ou em forma de resposta binária identificando quais observações são ou não são anomalias. Apesar da pontuação permitir a comparação entre as pontuações de duas observações diferentes, não fornece informações sobre quantas observações são consideradas reais anomalias na detecção.

Comumente, os algoritmos de detecção de anomalia utilizam o conceito de k-ésimo vizinhos mais próximos (comumente chamado de k-nearest neighbors (kNN)), muitas vezes este é um parâmetro a ser definido. Técnicas de detecção de anomalias são descritas nas próximas seções segundo as abordagens de densidade e distâncias.

3.1 Técnicas baseadas nas distâncias

O Algoritmo de detecção de anomalia que utiliza o kNN (KNN), proposto por Ramaswamy, Rastogi e Shim (2000), é um dos mais simples e utilizados algoritmos de ML. Ele utiliza a distância dos kNN para atribuir pontuações às observações. Dado um conjunto de dados com N observações, cada observação é agrupada aos seus k vizinhos mais próximos. Observações não atribuídas são consideradas anomalias. Há variantes do algoritmo que definem o primeiro ponto ou observação a iniciar e o número de estruturas a serem encontradas. Estas escolhas modificam os resultados alcançados.

Como exemplo, utilizamos o conjunto Iris (DUA; GRAFF, 2017), frequentemente referenciado em problemas de classificação, e definimos os k vizinhos mais próximos a serem identificados nas estruturas como 3. Comparando dois atributos (no caso, "sepal_length"e "sepal_width"), foram encontradas as anomalias na Figura 10, identificadas pelos pontos na cor vermelha.

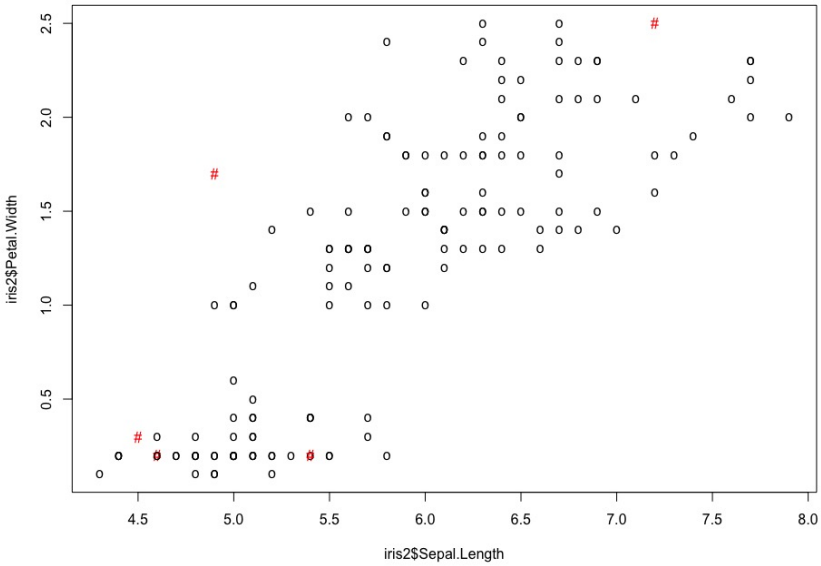


Figura 10 – Anomalias identificadas no conjunto de dados Iris com k=3

A partir desta conceituação, outras técnicas com base em uma matriz de distância foram desenvolvidas e técnicas baseadas em densidade também utilizam este conceito.

3.2 Técnicas baseadas na densidade

Os algoritmos baseados em densidade presumem que as observações que possuem um comportamento normal são as que estão em regiões com maior densidade segundo uma medida de comparação entre as observações. Dentre os algoritmos, o primeiro algoritmo de identificação de anomalias locais proposto e mais conhecido é o algoritmo LOF.

3.2.1 LOF

O algoritmo *Local Outlier Detection* (LOF), proposto por Breunig *et al.* (2000), compara a densidade local, atribuindo um fator de anormalidade local de cada observação por meio de um conjunto de kNN do ponto de teste com pontos dentro da vizinhança de cada membro do conjunto kNN. Sendo k um parâmetro a ser definido. Quanto maior a distância k do vizinho, maior a pontuação de anormalidade ou maior a discrepância do ponto em relação aos seus vizinhos.

Como exemplo, podemos identificar as 5 observações com maior índice de anormalidade segundo o algoritmo LOF com k=5 no conjunto de dados Iris (DUA; GRAFF, 2017) na Figura 11. Na figura, as anomalias identificadas são as observações na cor vermelha. Nota-se que, para esta identificação, deve-se escolher o corte, por exemplo uma quantidade de anomalias a serem identificadas, e a quantidade de k vizinhos mais próximos a percepção de normalidade. A representação 3D (por XQuartz [1]), auxilia a verificação e validação das anomalias de menor densidade no conjunto de dados para entendimento do resultado produzido pelo algoritmo.

A partir dos conceitos deste algoritmo, muitos outros foram criados, como Simplified-LOF, DBScan, COF, LoOp, entre outros.

3.3 Medidas de distância

Técnicas de detecção de anomalia baseada em densidade e distância utilizam a (dis)similaridade entre as observações. Ao comparar modelos de RB, as medidas utilizadas para a construção da matriz de dissimilaridade utilizada pelas técnicas são medidas de comparação entre grafos.

A medida encontrada para a comparação de redes Bayesianas é a distância Hamming proposta por Acid e Campos (2003), Tsamardinos, Brown e Aliferis (2006) e Perrier, Imoto e Miyano (2008). Essa distância compara as estruturas dos modelos e descreve o número de alterações que devem ser feitas em uma rede para que se transforme na rede com a qual está sendo comparada.

[1] <https://www.x.org/wiki/>

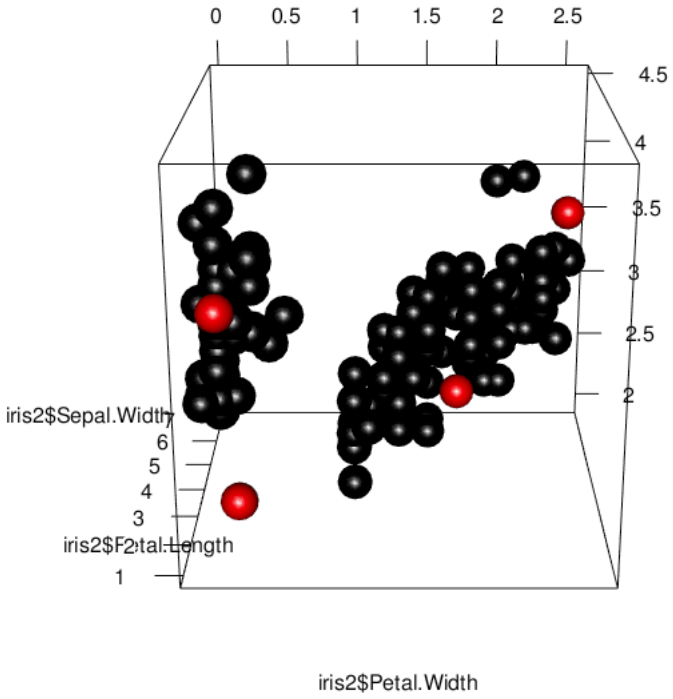


Figura 11 – Gráfico 3D do conjunto de dados Iris. Em vermelho, o resultado das 5 observações mais anômalas no conjunto de dados Iris identificados pela técnica LOF com k=5

3.4 Discussão

Apesar de extensivas discussões, para Chandola, Banerjee e Kumar (2009) os principais desafios ainda se concentram na adaptação aos diferentes domínios de aplicação. Alguns domínios importantes são a medida de similaridade escolhida, o tipo de dado, a ausência/presença de rótulos e o tipo de anomalia.

Com os algoritmos clássicos descritos, temos uma seleção de algoritmos que identificam anomalias pontuais e globais. Além disso, os algoritmos baseados na distância e densidade podem encontrar anomalias com diferentes contextos dependendo dos parâmetros de vizinhança e densidade definidos (SCHUBERT; ZIMEK; KRIEGEL, 2014; CAMPOS *et al.*, 2016). Desta forma, vários níveis de anomalias (locais e globais) podem ser encontrados dependendo dos critérios considerados. Isso é visto diretamente ao escolher o k no algoritmo KNN. No caso dos algoritmos baseados na densidade, o aumento de k engloba observações numa densidade menor em um mesmo grupo, identificando apenas possíveis anomalias globais e agregando as anomalias locais na estrutura.

Diferentes k influenciam nos tipos de anomalias encontradas, por isso deve ser considerada a possibilidade de avaliar o impacto da escolha deste parâmetro nas diferentes técnicas de detecção.

CAPÍTULO

4

PROPOSTA

"We've tended to forget that no computer will ever ask a new question."
(Grace Brewster Murray Hopper, 1906-1992)

A partir da lacuna sobre a partilha de dados apresentada na introdução, e da revisão teórica dos modelos representativos e de detecção de anomalias, é proposta a detecção de anomalia pela comparação de modelos representativos.

4.1 Metodologia

Pretende-se definir e descrever as variáveis selecionadas das entidades a serem comparadas juntamente aos especialistas dos dados. A proposta é que as entidades comparadas sejam instituições representadas pelos dados referentes a elas. A validação do preenchimento também deve ser acompanhada e validada com especialista da área.

Após seleção e validação do preenchimento dos dados, será realizada a análise sobre os valores faltantes de cada variável. Nesta análise, o cálculo da taxa de valores faltantes (*missings*) é realizado e serão identificados os mecanismos com que estes dados faltantes são relacionados a fim de indicar as técnicas para realizar a preenchimento de dados.

A seleção dos registros completos do conjunto real e os conjuntos de dados preenchidos serão comparados através de medidas estatísticas produzidas em cada conjunto de dados com o conjunto de dados reais sem preenchimento e ignorando os valores faltantes.

Os modelos recriados a partir dos dados que melhor representem os dados reais serão os modelos que recriarão os dados para este estudo. A validação dos modelos deve-se à capacidade de gerar dados em coerência com os dados reais. Estes novos conjuntos de dados recriados são

considerados como entidades para a comparação. Ou seja, para cada conjunto de dados recriados, são criados novos modelos sobre uma instituição recriada a partir de seus dados.

A comparação é realizada por meio de medidas de dissimilaridade entre redes Bayesianas já definidas na literatura. Comparando todos os modelos é construída uma matriz de dissimilaridade para utilização das técnicas de detecção de anomalia.

As anomalias identificadas pelas técnicas de detecção de anomalia serão validadas por especialistas nos dados por métricas e medidas escolhidas pelo especialista para comparação das entidades. Cada entidade identificada como anômala é analisada pelo especialista por meio de questionário apresentando as métricas e medidas escolhidas. Para cada entidade, é perguntado o nível de concordância que identifica, segundo a escala de Likert (1932), se a entidade apresentada é uma anomalia. A escala de concordância definida será: discordo plenamente, discordo em parte, nem concordo nem discordo, concordo em parte, concordo plenamente. Será também disponibilizado para cada pergunta um campo de notas para descrição do especialista, caso deseje acrescentar informação à análise.

4.2 Materiais e métodos

O primeiro passo é definir as técnicas de anonimização e preenchimento dos dados. Após este passo e a fim de entender quais os métodos que podem suprir e gerar conhecimento novo e útil, são escolhidos os métodos de representação dos dados (modelos) e as métricas de comparação destes modelos. Para a implementação das técnicas será utilizada a ferramenta R e os pacotes necessários para cada análise.

4.3 Anonimização dos dados

A anonimização dos dados acontecerá em três etapas, de forma manual (ou de casos específicos) e automatizada. Primeiramente é realizada a anonimização manual pelos especialistas dos dados, onde são identificados os registros de casos extremos ou reconhecidos pela sociedade. Após, os conjuntos de dados são anonimizados pelo sistema DiAS e análises automáticas.

O DiAS (2018) é um sistema web de pseudo-anonimização que garante a conformidade dos conjuntos de dados com o novo RGPD (Regulamento Geral de Proteção de Dados). Ele identifica datas, possíveis identificadores e mascara estes dados quando é necessário ou não permitido pelo comitê de ética que estes dados sejam utilizados na integridade. Por exemplo, datas completas são convertidas em mês e ano, se necessário. O sistema é reconhecido para anonimização de dados clínicos e por manter a utilidade dos dados para efeito de tratamento informacional e estatístico relevantes (DIAS, 2018).

Ao receber os dados, os valores de cada variável serão revalidados juntamente com especialista para que casos extremos ou de possível identificação não estejam presentes na análise ou sejam mascarados conforme indicação do especialista do contexto.

4.4 Falsos *missings*, mecanismos e preenchimento de dados faltantes

Os falsos *missings* são valores faltantes que podem ser preenchidos com o conhecimento dos valores de outras variáveis. A identificação deles é realizada juntamente a especialistas dos dados que identificam situações ou regras onde se pode realizar um preenchimento automático. Por exemplo, sempre que a variável A é 0, a variável B é 1. Chamamos isso de preenchimento determinístico.

Os valores faltantes que restarem, ou reais valores faltantes, são preenchidos por técnicas de preenchimento de dados. Definimos que serão utilizadas técnicas de preenchimento nas variáveis com taxa de *missing* de 10 a 90%. Para escolher a técnica de preenchimento são analisados os mecanismos com os quais os valores faltantes foram gerados. A identificação dos mecanismos foi proposta por RUBIN (1976). Segundo Rubin, cada dado faltante possui uma probabilidade de estar ausente. O motivo ou processo desta ausência, é chamado de mecanismo. Ele classifica os motivos dos valores faltantes em três: MCAR, MAR e MNAR.

Ao identificar que a probabilidade de ausência do dado é a mesma para todos os casos, define-se que os dados faltantes são ausentes completamente ao acaso, por isso o mecanismo designado é definido como *Missing completely at random* (MCAR) (BUUREN, 2018). São também definidos como MCAR os registros que não estão presentes na amostra, ou seja, todos os casos não constituintes de um estudo.

Se for identificado que a probabilidade da ausência do dado é a mesma apenas dentro de grupos definidos, é dito que os dados faltantes são ausentes aleatoriamente e o mecanismo dado para estes casos é o *Missing at random* (MAR). Apesar de semelhante, o MAR é mais geral que MCAR. MCAR ocorre quando a ausência do valor deve-se a "sorte" ou falta de explicação, já o MAR agrega alguma propriedade conhecida. Pode-se dizer, portanto, que MCAR é MAR em um cenário preestabelecido ou situação específica. Esses são os mecanismos mais comuns de dados faltantes identificados em métodos atuais, pois parte-se do pressuposto de conhecer o cenário de aplicação/análise.

Quado não é identificado MAR ou MCAR, é definido que o valor faltante é gerado por *Missing not at random* (MNAR). Estes são os casos mais complexos. Define-se que este valor faltante foi gerado pela falta da realização de análises em determinados cenários ou há uma causa

associada, porém não identificada.

Os mecanismos podem ser identificados por meio de análise exploratória com métodos estatísticos de análise de aleatoriedade dos *missings* aliado à análise visual. Na análise exploratória investiga-se se os dados estão totalmente omissos de forma aleatória ou são correlacionados a outras variáveis e preenchimentos. De acordo com as correlações, pode-se presumir o mecanismo correspondente com que os *missings* de cada variável foram gerados (LITTLE; RUBIN, 2019). Através dessa identificação aliada a análise de relação entre variáveis feita por especialistas nos dados e da área específica dos dados, podem ser presumidos os mecanismos para cada variável.

Serão utilizadas as correlações entre variáveis e a expertise do especialista para indicar se há relação entre variáveis do conjunto de dados e seus valores faltantes e, de acordo com a natureza dos dados em falta, pode-se sugerir um mecanismo que os descreve. O mecanismo MAR é sugerido quando os valores em falta estão relacionados com outros dados disponíveis no conjunto de dados (por exemplo, uma outra variável). O MNAR é sugerido quando os valores em falta estão relacionados com informação não observada, ou seja, não é identificável uma relação entres os valores faltantes e outras variáveis, e o comportamento também não é aleatório. Por fim, o MCAR existe quando não são encontradas quaisquer justificativas para os valores faltantes, ou seja, a sua existência assume um comportamento completamente aleatório.

As abordagens de técnicas de preenchimento utilizadas podem ser por preenchimento de moda ou média geral de cada variável ou por grupo, indicando que há relação entre variáveis. Outra técnica que será explorada é a técnica MICE (BUUREN; OUDSHOORN, 1999). O *Multivariate Imputation By Chained Equations*, ou apenas MICE, descrito por Buuren e Oudshoorn (1999), é um método de preenchimento múltiplo, usado para substituir valores de dados ausentes em um conjunto de dados de acordo com as presunções sobre o mecanismo da falta do dado. Neste estudo também será utilizado o pacote *MICE* (BUUREN *et al.*, 2015) da ferramenta R para o preenchimento. Espera-se que essa técnica obtenha um desempenho significativamente melhor por ser uma técnica robusta (PEREIRA *et al.*, 2020). Caso o desempenho, dos modelos ou da distribuição dos dados semelhantes aos dados originais, for semelhante à técnica mais simples de preenchimento, a técnica mais simples será escolhida para a execução nas instituições. Tais técnicas não geram anomalias pois elas transmitem e replicam o comportamento esperado dos dados presentes, seja pelas métricas de média e moda ou pelo mecanismo identificado.

4.5 Redes Bayesianas e estudo de caso

Para a criação de modelos representativos a partir dos dados foi escolhido os método Bayesiano. Como os dados são desconhecidos, eles podem evitar a tomada de decisões categóricas incorretas sobre as independências condicionais Neapolitan *et al.* (2004). Uma rede

Bayesiana é uma representação compacta de uma distribuição de probabilidade que geralmente é muito grande para ser tratada usando especificações tradicionais de probabilidade e estatísticas, como tabelas e equações (DARWICHE, 2010), possuindo uma maior explicabilidade do que outros tipos de modelos representativos. A partir delas é possível também sintetizar um grande volume de informação.

Neste contexto, redes Bayesianas foram eleitas para gerar e representar o modelo de comportamento de cada entidade. Em adição, o especialista nos dados ou área não específica de computação pode inferir evidência e informação, podendo gerar uma maior credibilidade ao modelo, representando o cenário real. A técnica possibilita esta inferência de maneira facilitada para não especialistas de diferentes áreas.

Vantagens do uso das redes Bayesianas:

- Transparência na análise de probabilidades e relação entre os dados;
- Método permite fazer uma modelação dos dados reais sem precisar deles após este processo;
- Agrega a análise estrutural, não só dos dados em si como a relação entre as variáveis do modelo;
- Fácil compreensão de especialistas que não possuam o cunho computacional em seus conhecimentos para entendimento do modelo gerado;
- Pode-se incluir evidência de diferentes fontes, agregando o conhecimento distribuído;
- Permite a replicação de dados artificiais no tamanho da base necessária para testes.

Para criar os modelos, será utilizada a ferramenta R e o pacote *bnlearn* (SCUTARI; SCUTARI; MMPC, 2020). Os modelos serão criados para todos os tipos de preenchimento de cada conjunto de dados.

A falta de preenchimento de dados é um desafio conhecido e preestabelecido ao lidar com dados clínicos. Serão abordados os diferentes mecanismos com que os dados faltantes são gerados e técnicas de preenchimento de dados.

Os modelos são construídos pelos dados totais recebidos das instituições, o que representa uma entidade neste estudo, produzindo um modelo para cada. A partir dos dados, são aprendidas as probabilidades de cada variável e construída a relação causal entre elas. As relações e probabilidades são calculadas utilizando as funções **discretize**, **hc**, **bn.fit** e a amostragem de dados a partir da rede criada pela função **rbn** do pacote de funções *bnlearn* descritas a seguir:

discretize(): Transforma os valores de cada variável torando-os mais adequados para a aprendizagem de estrutura.

hc: Aprende a estrutura de uma rede Bayesiana usando uma busca gulosa: *Hill-climbing* (HC) ou Tabu search (TABU).

bn.fit: Ajusta os parâmetros de uma rede Bayesiana atribuindo as probabilidades condicionais ao modelo.

rbn: Implementa uma amostragem direta. Os valores para os nodos raiz são amostrados a partir de sua distribuição (incondicional) e, em seguida, os de seus filhos condicionais nos respectivos conjuntos de pais. Isso é feito interativamente até que os valores tenham sido amostrados para todos os nós.

A partir dessas funções, a rotina básica adotada para criação de RB e dados artificiais a partir delas é a do Código 1.

Código-fonte 1 – Criação de redes Bayesianas e dados artificiais a partir de RB com funções do pacote *bnlearn* no R

```
1: #prepara os dados:
2: yads <- discretize(dados)
3: #gera a rede a partir dos dados:
4: entidade.net <- hc(yads)
5: #gera e associa as probabilidades de cada nodo:
6: entidade.fit <- bn.fit(entidade.net, yads)
7: novo_dataset <- rbn(entidade.fit, n=tamanho(dados))
```

4.5.1 Validação

A representatividade dos modelos será validada através da capacidade dos modelos gerarem dados similares aos dados reais. São extraídas características das variáveis dependendo do seu tipo: extraem-se moda das variáveis categóricas e a média, mínimo, máximo das variáveis quantitativas.

A comparação destas medidas dos conjuntos de dados dos registros completos selecionados e dos preenchidos com as medidas extraídas dos dados reais, ignorando os valores faltantes, definem a representatividade dos modelos.

4.6 Detecção de entidades anômalas

A comparação dos modelos das entidades permite quantificar a (dis)semelhança entre os mesmos. Através de uma matriz de dissemelhança é possível identificar casos raros ou anômalos que podem ser estudados e merecem uma análise posterior pelos especialistas nos dados ou entidades.

Nesta tese será realizada a detecção de anomalias por comparação de modelos representativos. Essa é uma nova abordagem, onde será utilizada a medida de dissimilaridade entre grafos (distância estrutural de Hamming) para a construção da matriz de dissimilaridade entre as entidades e, a partir desta, será realizada a detecção da entidade considerada anômala pelas técnicas.

4.7 Validação

A validação das anomalias identificadas é feita por especialistas na área do caso de estudo. Será apresentado um questionário com informações dos dados referentes ao(s) modelo(s) da(s) entidade(s) identificadas como anômalas. Sobre cada identificação deve-se responder na escala proposta por Likert (1932) o quão concordante é com a afirmação “esta entidade é anômala”.

CAPÍTULO

5

CASO DE ESTUDO: DADOS CLÍNICOS

"*We should be suspicious of any dataset (large or small) which appears perfect.*"

(David Hand)

O caso de estudo pertence ao cenário obstétrico de Portugal. O problema foi identificado no contexto da preocupação mundial sobre a crescente taxa de cesarianas realizadas nos hospitais.

5.1 Contextualização

A cesariana apresenta um risco maior de apresentar complicações no parto do que o parto normal, também referenciado como parto eutócico ou eutócito em Portugal. A grávida que realiza a cesariana está exposta aos riscos de anestesia, complicações como infecções, grande perda de sangue, coágulo de sangue nas pernas e pulmões Mylonas e Friese (2015). O período de recuperação do parto cesárea é muito maior do que após um parto vaginal, o que adiciona a exposição maior a problemas e infecções hospitalares. Tais riscos são refletidos na taxa de mortalidade: conforme a taxa de cesariana diminui nos países, a morbidade e mortalidade neonatal também diminui (KOTASKA, 2015).

Em um cenário ideal, os partos por cesarianas são realizados apenas em situações em que a grávida ou o bebê estão em situação de risco de vida. No entanto, o cenário real parte da utilização em partos de alto risco para uma possível banalização deste tipo de parto e preocupação de seus elevados índices. Para fins de entendimento dos índices de cesáreas e classificação do tipo de parto, a OMS (ORGANIZATION *et al.*, 2015) conclui através de uma revisão sistemática que a forma de classificar que melhor cumpre os requisitos estabelecidos pela organização é a

baseada nas características da gravidez, indicando o uso da classificação de Robson (ROBSON, 2001) para este fim.

A classificação de Robson possui quatro critérios e, a partir deles, todas as gestantes podem ser classificadas em 10 grupos mutuamente exclusivos e totalmente inclusivos. O critérios são sobre os antecedentes obstétricos, o tipo da gestação, a forma como se desencadeia o parto e a idade gestacional em que o parto ocorre. A partir destes critérios, as gestantes são classificadas em um dos grupos expostos na Tabela 3. Os critérios são simples e facilmente adquiridos por médicos ou enfermeiros. Os grupos são mutuamente exclusivos.

Tabela 3 – Descrição dos grupos da classificação de Robson

Grupo	**Descrição**
1	Nulípara, único cefálico $\geq$ 37 sem, TP espontâneo
2	Nulípara, único cefálico $\geq$ 37 sem, TP induzido OU Ces antes TP
3	Multípara sem cesariana anterior, único cefálico $\geq$ 37 sem, TP espontâneo
4	Multípara sem cesariana anterior, único cefálico $\geq$ 37 sem, ind OU Ces antes TP
5	Multípara com 1 OU mais Ces ant, único cefálico $\geq$ 37 sem
6	Todas nulíparas com apresentação pélvica
7	Todas multíparas com apresentação pélvica (incluindo com cesariana anterior)
8	Todas gestações múltiplas (incluindo com cesariana anterior)
9	Todas córmicas ou oblíquas (incluindo com cesariana anterior)
10	Todas único cefálico < 37 sem (incluindo com cesariana anterior)

*TP = Trabalho de parto
*sem = semanas de gestação
*Ces = cesariana

Ao utilizar a classificação de Robson, tem-se a descrição nos grupos e neles pode-se interpretar como partos cesáreos evitáveis, não evitáveis ou pouco evitáveis (ROBSON, 2001). Apesar da classificação, as razões da crescente realização de partos cesáreos em grupos em que este seria evitável são desconhecidas e os hospitais muitas vezes não compartilham seus dados integralmente para uma análise pormenorizada.

Ao compartilhar dados entre instituições pode-se melhorar as descobertas e induzir à generalização de convicções, crenças e ideias que são locais para decisões globais, por meio de amostras maiores. Porém, atualmente a partilha desses dados requer alta confidencialidade e por isso existem muitas barreiras burocráticas gerais do uso destes dados e burocracias de cada instituição. Uma forma de recriar os dados reais com representatividade significativa pode ser a solução nestes casos. As redes Bayesianas (RB) podem recriar novas amostras replicando as características dos dados, agregada a construção da estrutura do modelo representativo. Os modelos de RB e sua construção oferecem a capacidade de descrever entidades que representam as seções de obstetrícia em hospitais. Com isso, almeja-se descrever os registros obstétricos e, consequentemente, a análise de outras variáveis que podem estar também relacionadas aos tipos de partos podem ser analisadas.

Nesse contexto, o objetivo é criar modelos representativos com variáveis que definem ou

possam interferir no tipo de parto para cada hospital juntamente com especialistas nos dados. As redes Bayesianas servem para descrever o tipo de decisão da seção de obstetrícia. Essas redes serão criadas localmente e os dados não serão compartilhados. A partir dos modelos construídos é possível comparar esses dados e detectar os modelos que diferem substancialmente dos outros. Esses modelos identificados são os definidos como anomalias e representam entidades ou instituições com comportamento anômalo.

Para realizar este estudo no contexto de dados médicos e clínicos em Portugal o acesso aos dados é permitido através do pedido ao RAI (responsável pelo acesso à informação) em algumas instituições e à comissão de ética, obrigatório em todas instituições. É um processo longo pela percepção das variáveis possíveis de incluir no estudo, além dos dados requeridos serem especificados por período e razão, e utilizados para um caso de estudo também específico. Para além das regras gerais, cada instituição enviou os procedimentos padrões da instituição, como documentos de responsabilidade pelos dados e indicação de responsáveis de cada instituição, um elo da pesquisa com a instituição a qual foram requeridos os dados.

Após a criação dos modelos de cada entidade, os dados que serão utilizados serão os recriados pelo modelo, não os originais. Os dados recriados não conterão informações particulares que necessitem de autorização e terão, assim, o acesso disponibilizado.

A metodologia é apresentada através de etapas que delimitam e norteiam o desenvolvimento sobre o estudo dos métodos e do caso de estudo.

5.2 Metodologia

No estudo da detecção de anomalias em dados de seções obstétricas de hospitais estamos interessados em descobrir os motivos pelos quais sua instituição teve um melhor ou pior comportamento, as influências de atributos e quais são as diferenças entre outras instituições.

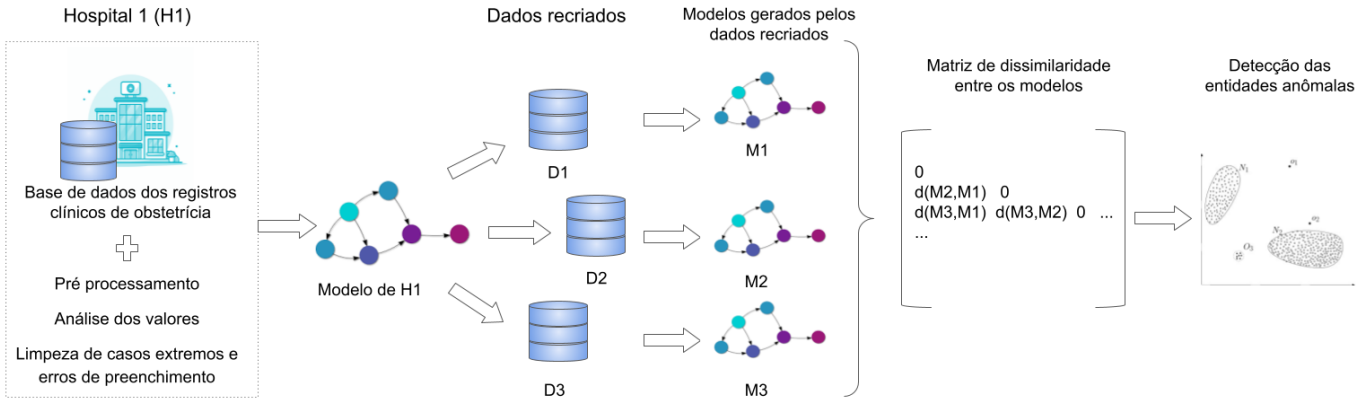
Para a avaliação de dados e informação médica, os dados foram cedidos pela empresa VirtualCare [1] após o acesso autorizado pelos comitês de ética. A empresa possui um software denominado Obscare (CRUZ-CORREIA; AMORIM, 2008) que recolhe informação sobre partos em treze hospitais de Portugal. Foram requeridos os dados de 9 hospitais e, até o momento, obtivemos resposta de três hospitais que são participantes deste caso de estudo, com as autorizações e pareceres dos respectivos comitês de ética: HSO (85/2020), CHEDV (CA-371/2020), CHVNGE (192/2020) que serão depois melhor descritos na Sessão 5.3. Cada registro corresponde a uma gestação e parto e as informações são sobre antecedentes clínicos, gestação e outras informações relevantes até o momento do parto. Em caso de gêmeos, o tipo de parto do segundo bebê é

[1] Saber mais em http://virtualcare.pt/

desconsiderado para manter um registro apenas pela gestação. Porém a informação sobre a quantidade de recém-nascidos está contida na base de dados, portanto, sabe-se se a gestação é gemelar ou não.

As redes Bayesianas são construídas pelos dados reais de cada instituição. Os modelos de RB de cada hospital ou seção obstétrica de hospitais gerados a partir dos dados reais são utilizados para a amostragem de dados. A partir desses dados recriados são gerados modelos representativos. Estes modelos criados pelos dados recriados são comparados pela distância Hamming. O processo é ilustrado na Figura 12.

Figura 12 – Processo de criação de modelos e amostragem de dados deles em uma instituição exemplo H1



A comparação de modelos reflete o quão diferentes são as instituições entre si e permite posteriormente analisar, de acordo com as entidades identificadas, se essas diferenças são positivas ou negativas na análise proposta. Por meio desse comparativo algumas instituições podem ser consideradas instituições anômalas. Anomalia significa que é muito diferente das outras instituições (neste caso) que tenham comportamento diferente consideradas positivas (por exemplo, menor índice de partos cesáreos, maioria dos partos supervisionados, entre outros) ou negativas (menor índice de partos normais, maior quantidade de complicações, entre outros). Ou seja, as entidades representadas pelos modelos podem ser destoantes das restantes por serem ideais a serem seguidos ou, pelo contrário, por terem um comportamento regular.

O caso de estudo também agrega a análise de *missings* e a relação entre as variáveis. Em primeiro lugar, analisamos os valores ausentes. Variáveis com mais de 90% de dados faltantes foram retiradas da amostra. Embora as RB possam lidar com itens de dados ausentes, escolhemos usar métodos de preenchimento de dados indicados na identificação de mecanismos para os tipos de *missings* identificados em cada variável. As variáveis são selecionadas conforme a sua significância estatística calculada pelo $p - value$ entre a variável analisada e variável tipo de

parto. A análise da significância de cada variável com o tipo de parto realizado auxilia a proposta de novas variáveis à classificação e pode servir de apoio para a decisão do especialista.

A validação será realizada pela aplicação de questionário apresentando a distribuição dos grupos de Robson pelos tipos de parto e taxas dos tipos de parto. O questionário será respondido por um especialista obstetra. No contexto em que este trabalho foi desenvolvido juntamente ao período histórico de uma pandemia por coronavírus, a validação foi diminuída a uma pessoa que se mostrou disponível. A especialista obstetra realizará a validação apenas das anomalias encontradas comparando com seu histórico e conhecimento das seções obstétricas de Portugal. A especialista está inserida no contexto obstétrico com vasta experiência e não é vinculada aos hospitais aderentes deste estudo.

5.3 Base de dados

As variáveis utilizadas neste caso de estudo são sobre gestação e partos de grávidas em hospitais portugueses. Serão analisados dados cedidos por três hospitais do Norte de Portugal[2]: Hospital Senhora de Oliveira de Guimarães (HSOG) autorizado pela Comissão de Ética pela Saúde (CES) na referência 85/2020, Centro Hospitalar Vila Nova de Gaia/Espinho (CHVNGE) autorizado pela CES na referência 192/2020 e Centro Hospitalar de Entre Douro e Vouga de Santa Maria da Feira (CHEDV) autorizado pela referência 371/2020, que utilizam o sistema Obscare [3] para a coleta e armazenamento de dados. Os hospitais serão chamados de H1, H2 e H3. Especialistas nos dados revisaram a descrição e cederam a versão de testes do sistema que realiza a coleta de dados dos três hospitais para revisão e manipulação da ferramenta.

Foram recolhidos pelo Obscare os dados do período da data de implantação a dezembro de 2020. A busca destes dados em H1 (a partir de 01/2015) retornou 11.705 registros; em H2 (a partir de 03/2015) retornou 7.743; em H3 (a partir de 10/2016) retornou 6.572 registros. Inicialmente foram recolhidas 90 variáveis e estas foram filtradas novamente conforme taxa de preenchimento após validação dos valores preenchidos (Tabela 5) e análise dos especialistas.

5.3.1 Descrição das variáveis

A escolha das variáveis foi realizada juntamente a dois especialistas nos dados e ferramenta aplicados. Foram utilizadas variáveis referentes à gravidez, à admissão, ao parto e às possíveis variáveis relacionadas ao tipo de parto, como antecedentes obstétricos e pessoais da gestante, garantindo que as variáveis relativas a classificação de Robson (ROBSON, 2001) compusessem a seleção das variáveis. As variáveis recebidas são:

[2] Autorizações dos comitês de ética de cada hospital no Anexo A
[3] Descrição completa sobre a ferramenta em http://virtualcare.pt/portfolio/vc-obscare/

- A_GESTA: Número de gestações que a grávida teve até o momento da gravidez (gravidez atual, partos e abortamentos)
- ALCOOL: Indica se a grávida ingeriu álcool durante a gestação
- APRESENTACAO_24 - 42: Apresentação do bebê na 24ª a 42ª semana de gestação
- APRESENTACAO_ADMISSAO: Posição do feto na primeira verificação no hospital (posição cefálica, pélvica etc.)
- BACIA: avaliação clínica da pelve óssea (bacia) para saída do feto
- BISHOP_SCORE: Pontuação da escala de maturação cervical (BISHOP) utilizada para prever se a indução ou condução do trabalho de parto será necessária
- BISHOP_CONSISTENCIA: Na classificação de Bishop a consistência pode ser classificada em 3 estados: Duro (0), Médio (1) ou Macio (2)
- BISHOP_DESCIDA: Medida em centímetros, scores em: 0 se for >3 cm, 1 se >2 cm, 2 se >1 cm, 3 se >0 cm
- BISHOP_DILATACAO: Na classificação de Bishop, a dilatação pode ser classificada em 4 estados: Fechado (0), 1-2 cm (1), 3-4 cm (2), ou 5+cm (3)
- BISHOP_EXTINCAO: Índice utilizado na classificação Bishop
- BISHOP_POSICAO: Índice utilizado na classificação Bishop: Posterior (0), Intermediário (1), Central (2)
- CARDIACA: Diagnóstico de doenças cardíacas pré-gravidez (doença reconhecida antes da gravidez atual)
- CEREBRAL: Diagnóstico de doenças cerebrais pré-gravidez (doença reconhecida antes da gravidez atual)
- CESARIANAS_ANTERIOR: Quantidade de partos cesáreas anteriores
- CESARIANAS_MOTIVO_ANTERIOR: Descrição do motivo de partos anteriores feitos por cesariana
- CIGARROS: Indica se a grávida fumou durante a gestação
- DATA_PARTO: Data do início do parto
- DIABETES_GESTACIONAL: Diagnóstico de diabetes na gravidez (doença devido à gravidez ou diagnosticada com a gravidez)

- DIABETES_GESTACIONAL_DIETA: Diagnóstico de diabetes acompanhada de dieta específica na gravidez (doença devido à gravidez ou diagnosticada com a gravidez)
- DIABETES_GESTACIONAL_INSULINA: Diagnóstico de diabetes na gravidez acompanhada com a aplicação de insulina na gravidez (doença devido à gravidez ou diagnosticada com a gravidez)
- DIABETES_GESTACIONAL_ANTIBIO: Diagnóstico de diabetes na gravidez acompanhada de antibiótico (doença devido à gravidez ou diagnosticada com a gravidez)
- DIABETES_MATERNA: Diagnóstico de diabetes prévia a gravidez (doença reconhecida antes da gravidez atual)
- DIABETES_TIPO1: Descrição de diabetes, se tipo 1, com diagnóstico pré-gravidez (doença reconhecida antes da gravidez atual). Tipo de diabetes materna. Se variável DIABETES_MATERNA está preenchida com sim, esta ou DIABETES_TIPO2 devem estar preenchidas. Especificação de DIABETES_MATERNA
- DIABETES_TIPO2: Descrição de diabetes, se tipo 2, com diagnóstico pré-gravidez (doença reconhecida antes da gravidez atual). Tipo de diabetes materna. Se variável DIABETES_MATERNA está preenchida com sim, esta ou DIABETES_TIPO1 devem estar preenchidas. Especificação de DIABETES_MATERNA
- E_ALT_UT: Altura uterina. Medição da altura/tamanho da barriga em centímetros.
- ESTIMATIVA_PESO_ECO_24 - 42: Estimativa de peso do bebê na 24ª a 42ª semana de gestação feita pelo exame de ecografia
- ESTUPEFACIENTES: Indica se a grávida utilizou estupefacientes (entorpecentes) durante a gestação
- EUTOCITO_ANTERIOR: Quantidade de partos eutócitos anteriores, feitos via vaginal sem uso de nada
- FORCEPS_ANTERIOR: Quantidade de partos eutócitos anteriores, feitos via vaginal com uso de forcéps
- GRUPO_ROBSON: Grupo de Robson associado de acordo com a classificação de Robson
- GS: Grupo sanguíneo e RH da grávida
- HEMATOLOGICA: Diagnóstico de doenças hematológicas pré-gravidez (doença reconhecida antes da gravidez atual)

- HIPERTENSAO_CRONICA: Diagnóstico de hipertensão crônica pré-gravidez (doença reconhecida antes da gravidez atual)
- HIPERTENSAO_GESTACIONAL: Diagnóstico de hipertensão na gravidez (doença devido à gravidez ou diagnosticada com a gravidez)
- HIPERTENSAO_PRE_ECLAMPSIA: Diagnóstico de hipertensão pré-eclâmpsia na gravidez (doença devido à gravidez ou diagnosticada com a gravidez)
- IDADE_GESTACIONAL_ADMISSAO: Semanas de gestação no momento da admissão, quando nasceu o bebê
- IDENTIFICADOR: ID anonimizado
- IDADE_MATERNA: Idade da grávida no momento do parto
- IMC: Índice de massa corporal da grávida no início da gravidez
- NUMERO_CONSULTAS_PRE_NATAL: Nº de consultas pré-natais
- PESO_ADMISSAO_INTERNAMENTO: Peso da grávida no final na gravidez, quando é internada para ter o bebê
- PESO_INICIAL: Peso da grávida no início da gravidez
- RESPIRATORIA: Diagnóstico de doenças respiratórias pré-gravidez (doença reconhecida antes da gravidez atual)
- RPM: Indica se houve rotura prematura de membranas e/ou rotura da bolsa amniótica antes do início do trabalho de parto
- TIPO_GRAVIDEZ: Tipo de gravidez atual, se foi espontânea, se usou
- TIPO_PARTO: Tipo de parto realizado da gravidez atual inseminação, FIV, entre outras
- TRABALHO_PARTO_ESPONTANEO: Indica se o trabalho de parto foi espontâneo
- TRABALHO_PARTO_INDUZIDO: Indica se o trabalho de parto foi induzido
- VENTOSAS_ANTERIOR: Quantidade de partos eutócicos anteriores, feitos via vaginal com auxílio de ventosas
- VIGIADA: Indica se a gravidez foi vigiada (mais de 5 consultas para acompanhamento da gestação realizadas)
- VIGIADA_HOSPITAL: Indica se a gravidez foi vigiada em hospital

- VIGIADA_PARICULAR: Indica se a gravidez foi vigiada com plano particular
- VIGIADA_CENTRO_SAUDE: Indica se a gravidez foi vigiada em centro de saúde
- VIGIADA_NESTE_HOSPITAL: Indica se a gravidez foi vigiada no mesmo hospital do parto

Na seleção de variáveis foi requerida em especial a variável sobre a supervisão e acompanhamento da gestação (variável "VIGIADA"). Esta variável pode ser relevante e incentivar na escolha do tipo de parto e também foi incentivada a sua análise agregando a classificação de Robson por Ray e Jose (2017). Vê-se que há variáveis sobre entrada/admissão da grávida e sobre a hora do parto. Foram recolhidas pois podem existir valores diferentes nos dois contextos. Como exemplo, em relação ao trabalho de parto, a grávida pode chegar com trabalho de parto espontâneo e na hora os médicos resolverem induzir. Para a classificação dos grupos de Robson, as variáveis são sobre a hora do parto.

5.4 Anonimização dos dados

O processo passou por duas análises. A primeira pelo sistema DiaS[4] e a segunda uma análise exploratória juntamente ao especialista dos dados. O sistema DiaS retirou identificadores e transformou datas para o formato MM/YYYY, retirando o dia do parto. Com o especialista, foram analisados os casos raros e todos foram aprovados para conter na base de dados, pois os dados reais são utilizados apenas para a fase de construção da rede Bayesiana.

5.5 Pré-processamento dos dados

Primeiramente foram retiradas variáveis que não seriam aproveitadas no estudo, como estimativa de peso das ecografias semanais e posição do feto da 24^a a 42ª semana. Estas variáveis não foram consideradas por estarem vazias ou com preenchimento menores a 1% do total de registros. Foram também excluídos registros duplicados de uma mesma gestação. A informação sobre a quantidade de bebês da gestação foi mantida para percepção sobre a gestação, se é gemelar ou não. O tipo de parto dos registros da gestação gemelar é o do primeiro recém-nascido. Interessa à pesquisa os dados sobre o percurso da gestação, sobre a grávida e sobre o tipo de parto, por isso essas etapas foram definidas. Após estes processos, foi realizado o primeiro cálculo de valores faltantes por variável.

[4] Descrição completa em http://hltsys.pt/dias-project-en/

5.5.1 Validação dos valores e preenchimento dos falsos missings

Com base na classificação de Robson e nas características de cada grupo, fez-se uma reclassificação dos registros classificados erroneamente. Esta reclassificação foi seguida de um preenchimento determinístico de informação, por meio de conceitos definidos por especialistas e conclusões tiradas dos próprios dados sobre as variáveis. A este processo chamamos de preenchimento de falsos valores faltantes e foi sugerido que este preenchimento fosse automático para os sistemas de coleta de dados (GELATTI; RODRIGUES; CORREIA, 2020). As definições foram revisadas por especialista da área e pelo especialista da coleta dos dados cedidos. Após esta validação, foram eliminados os registros duplicados, mantendo um registro por gestação.

Após a reclassificação e preenchimento determinístico de falsos valores faltantes, foi realizada uma análise por variável juntamente com especialista nos dados. Esta análise teve por objetivo encontrar inconsistência nos dados e erro de preenchimento das variáveis. Ao analisar as variáveis referentes às semanas de gestação na chegada ao hospital e semanas de gestação na hora do parto, foram encontrados valores superiores a 48 semanas (2 registros de 48 e 1 registro de 58 semanas). Por ordem médica, as gestações são geralmente interrompidas às 42 semanas sugerindo a cesariana a grávidas que não entraram em trabalho de parto até este período. Por esta razão, definiu-se que valores encontrados além de 45 semanas foram considerados erro de preenchimento nos dados e foram preenchidos com valor nulo.

Nos valores preenchidos sobre a apresentação do bebê foram encontradas a posição específica: cefálica de vértice, cefálica defletida, pélvica, face, oblíqua e espádua; e a situação da posição: transversa e longitudinal. Todos valores referentes a situação transversa (valores "transversa", espádua e oblíqua) foram transformados para a situação transversa. Os outros valores foram agrupados quanto a posição cefálica (face, cefálica de vértice) e pélvica. Situação parecida ocorre quanto aos tipos de parto. Esses foram reagrupados pelos tipos cesariana, partos auxiliados (ou instrumentados) e partos eutócicos (partos normais).

Após estes pré-processamentos, podemos comparar a taxa de valores faltantes nos dados depois do preenchimento dos falsos *missings* e preenchimento determinístico por hospital. As variáveis com a taxa de valores faltantes superior a 90% após o pré-processamento foram excluídas do estudo, pois concluímos que é uma quantidade pequena para o preenchimento basear-se, já que não há variáveis relacionadas com o preenchimento. São essas as variáveis excluídas: PESO_ADMISSAO_INTERNAMENTO, CESARIANAS_MOTIVO_ANTERIOR, E_ALT_UT, BACIA, BISHOP_SCORE, BISHOP_CONSISTENCIA, BISHOP_DESCIDA, BISHOP_DILATACAO, BISHOP_EXTINCAO, BISHOP_POSICAO. Também foi excluída a variável "G_TERAPEUTICA"por conter descrição de medicamentos. Foi escolhido não utilizar essa variável nesta análise para garantir maior anonimização dos dados e por não utilizarmos qualquer mineração de texto na análise.

Tabela 4 – Referência sobre a descrição dos grupos da classificação de Robson

Grupo	Descrição
1	Nulípara, único cefálico $\geq$ 37 sem, TP espontâneo
2	Nulípara, único cefálico $\geq$ 37 sem, TP induzido OU Ces antes TP
3	Multípara sem cesariana anterior, único cefálico $\geq$ 37 sem, TP espontâneo
4	Multípara sem cesariana anterior, único cefálico $\geq$ 37 sem, ind OU Ces antes TP
5	Multípara com 1 OU mais Ces ant, único cefálico $\geq$ 37 sem
6	Todas nulíparas com apresentação pélvica
7	Todas multíparas com apresentação pélvica (incluindo com cesariana anterior)
8	Todas gestações múltiplas (incluindo com cesariana anterior)
9	Todas córmicas ou oblíquas (incluindo com cesariana anterior)
10	Todas único cefálico < 37 sem (incluindo com cesariana anterior)

*Trabalho de parto (TP)
*sem = semanas de gestação
*Ces = cesariana *ant = anterior

As principais validações e preenchimentos realizados foram pelo preenchimento determinístico baseado nos grupos de Robson. Esta classificação, com a descrição novamente apresentada na Tabela 4, agrega informações relativas a outras variáveis dos registros, como cesarianas anteriores, apresentação fetal, entre outras já relatadas. De acordo com a quantidade de alterações e reclassificações em cada grupo, as situações mais frequentes encontradas foram:

Grupo 6:

registros não preenchidos em *A_PARA* (todos devem ser primeiro filho, *A_PARA* = 0), foram preenchidos com 0

Grupo 5:

registros sem preenchimento em cesarianas anteriores (esperado: com cesariana)

registros sem filhos anteriores, todos outros com *A_PARA* >0

registros com trabalho de parto espontâneo (reclassificados para o grupo 3)

registros com ausência de trabalho de parto e tipo de parto cesariana reclassificados para o grupo 4

registros com trabalho de parto induzido reclassificados para o grupo 4

Grupo 2:

registros com trabalho de parto espontâneo entre outras características do grupo 1 foram reclassificados para o respectivo grupo

registros com cesariana anterior foram reclassificados para o grupo 5

registros não preenchidos com o número de cesarianas anterior foram preenchidos com zero

Grupo 4:

registros com cesarianas anteriores a gestação (neste grupo o esperado é 0), características do grupo 1, foram reclassificados para o respectivo grupo

registros com trabalho de parto espontâneo

registro de ausência de trabalho de parto, porém não cesariana (com parto eutócico cefálico), o que é um erro. Foi classificado como indefinido (definido como o estado "U", *undefined*)

As situações raramente encontradas foram nos grupos:

Grupo 1:

registros com cesariana anterior, reclassificados para o grupo 5

registros com falta de preenchimento sobre cesariana anterior (no caso deste grupo, igual a zero) foram preenchidos

registros com falta de preenchimento de filhos anteriores (variável *A_PARA*) foram preenchidos com 0

Grupo 3:

registros com cesariana anterior, reclassificados no grupo 5

Grupo 8:

registros com gestação única foram reclassificados para os respectivos grupos (grupo 3 e grupo 4)

Grupo 10:

registros que possuíam "-1" em semanas de gestação. Este grupo é composto com gestação inferior a 37 semanas. Por acreditar que este foi um erro na classificação automática, estes registros foram reclassificados como grupo de Robson indefinido (U)

Referente ao Grupo 8, grupo onde se encontram apenas gestações de gemelares, os registros são classificados pelo sistema em determinada hora (não especificada) que, se por algum motivo, na hora do cálculo houver mais de um registro ativo sobre a gestação, é colocado no grupo 8 pois é considerado como mais de um recém-nascido.

Foram retirados os registros duplicados, mantendo 1 registro por gestação, resultando no que é considerado para as próximas análises como conjunto de dados reais. Na Tabela 5 vemos as variáveis e os valores totais de taxas de *missings* antes e após o pré-processamento e preenchimento determinístico.

Nota-se que o preenchimento das variáveis (e a falta dele) segue como um padrão nos hospitais. Apesar deste caso de estudo conter uma amostra dos hospitais do norte de Portugal, acredita-se que a falta de preenchimento seja característica comum aos hospitais portugueses pois acontece em todas bases de dados dos hospitais recebidos.

Alguns dados em discordância foram encontrados na variável IMC e tratados como ruído, como o registro com o maior peso (169kg) possui um IMC de 57,8. Os valores de IMC acima deste eram *[70 , 86.4 , 94.5, 146.5, 172.2, 187.8, 202.9, 307.5, 318.3, 448.8]* Ao refazer o cálculo para encontrar as alturas referentes a estes valores foram encontrados valores menores a $1m$. O valor de IMC desses 10 registros com altos valores de IMC foram retirados por se considerar erro, sendo validado pelo especialista. Outra variável é o N^{o} de gestações. Ao listar os valores, havia valores com o valor zero. A variável é sobre a quantidade de gestações, incluindo a gestação atual e abortos. Visto que os dados são somente de pessoas com gestação, os dados apresentados em zero foram corrigidos no passo a seguir para o valor um.

Tabela 5 – Comparativo da taxa de *missings* por variável pré e após preenchimento determinístico

Hospitais	H1		H2		H3	
Total (n)	11.705		7.743		6.572	
Taxa de missings	pré(%)	pós(%)	pré(%)	pós(%)	pré(%)	pós(%)
DATA_PARTO	0	0	0	0	0	0
IDADE_MATERNA	0	0	0	0	0	0
GS	0	0,67	0	0,25	0	0,06
PESO_INICIAL	40,8	40,8	15,83	15,83	7,76	7,76
PESO_ADMISSAO_INTERNAMENTO	95,3	95,3	85,38	85,38	96,71	96,71
IMC	41,73	41,73	20,29	20,29	8,31	8,31
CIGARROS	91,16	0	86,7	0	84,25	0
ALCOOL	98,27	0	97,15	0	99,09	0
ESTUPEFACIENTES	99,93	0	99,85	0	99,6	0
A_PARA	38,97	9,48	40,76	8,61	47,5	10,96
A_GESTA	0,02	0,02	0,01	0,01	0	0
EUTOCITO_ANTERIOR	59,03	59,03	59,4	59,4	73,11	73,11
VENTOSAS_ANTERIOR	79,03	79,03	79,9	79,9	90,26	90,26
FORCEPS_ANTERIOR	82,21	82,21	83,78	83,78	99,18	99,18
CESARIANAS_ANTERIOR	70,18	14,82	73,49	14,04	88,21	19,05
CESARIANAS_MOTIVO_ANTERIOR	100	100	100	100	100	100
TIPO_GRAVIDEZ	9,25	9,25	9,01	9,01	32,36	32,36
VIGIADA	0,05	0,05	0,05	0,05	0,02	0,02
NUMERO_CONSULTAS_PRE_NATAL	37,01	37,01	22,14	22,14	11,82	11,82
VIGIADA_HOSPITAL	98,74	61,14	97,66	24,34	95,62	12,99
VIGIADA_PARICULAR	67,92	67,92	83,7	83,7	84,53	84,53
VIGIADA_CENTRO_SAUDE	53,44	53,44	46,8	46,8	34,95	34,95
VIGIADA_NESTE_HOSPITAL	61,8	61,8	25,36	25,36	14,67	14,67
NUM_RN	0	0	0	0	0	0
E_ALT_UT	99,99	99,99	94,23	94,23	100	100
BACIA	89,31	89,31	76,61	76,61	95,24	95,24
APRESENTACAO_ADMISSAO	0,98	0	1,45	0	7,61	0
BISHOP_SCORE	99,21	99,21	79,16	79,16	98,45	98,45
BISHOP_CONSISTENCIA	99,49	99,49	83,84	83,84	98,54	98,54
BISHOP_DESCIDA	99,51	99,51	83,84	83,84	98,63	98,63
BISHOP_DILATACAO	99,48	99,48	83,55	83,55	98,49	98,49
BISHOP_EXTINCAO	99,49	99,49	83,7	83,7	98,54	98,54
BISHOP_POSICAO	99,5	99,5	83,8	83,8	98,57	98,57
IDADE_GESTACIONAL_ADMISSAO	1,36	1,36	0,54	0,54	0,72	0,72
TRAB_PARTO_ENTRADA_ESPONTANEO	64,42	0	73,95	0	66,98	0
TRAB_PARTO_ENTRADA_INDUZIDO	91,86	0	85,33	0	97,64	0
RPM	85,1	0	93,06	0	90,2	0
HIPERTENSAO_CRONICA	97,9	0	97,64	0	96,93	0
HIPERTENSAO_GESTACIONAL	98,89	0	97,8	0	98,78	0
HIPERTENSAO_PRE_ECLAMPSIA	98,19	0	98,59	0	98,39	0
DIABETES_GESTACIONAL	91,99	0	90,22	0	87,43	0
DIABETES_GESTACIONAL_DIETA	96,68	0	93,27	0	91,01	0
DIABETES_GESTACIONAL_INSULINA	98,8	0	98,64	0	98,6	0
DIABETES_GESTACIONAL_ANTIBIO	98,62	0	98,53	0	98,36	0
DIABETES_MATERNA	99,68	0	99,63	0	99,47	0
DIABETES_TIPO1	99,92	0	99,87	0	99,77	0
DIABETES_TIPO2	99,84	0	99,88	0	99,83	0
HEMATOLOGICA	97,2	0	95,84	0	94,55	0
RESPIRATORIA	97,27	0	95,53	0	93,99	0
CEREBRAL	99,1	0	98,79	0	98,63	0
CARDIACA	94,41	0	93,49	0	93,26	0
TIPO_PARTO	0	0	0	0	0	0
APRESENTACAO_NO_PARTO	0,11	0	0,01	0	0,12	0
TRAB_PARTO_NO_PARTO	0	0	0	0	0	0
SEMANAS_GESTACAO_PARTO	0,01	0,19	0	0,13	0,03	0,18
GRUPO_ROBSON	0,08	0	6,12	0	0,15	0

5.6 Preenchimento de dados

Ao preencher dados em valores faltantes em cenários reais, não há a possibilidade de validação conferindo a referência do dado preenchido na tabela verdadeira. A validação neste caso parte do princípio que se tem um conjunto de dados completos onde se pode gerar artificialmente os valores faltantes. No cenário apresentado neste caso de estudo será avaliado o impacto do preenchimento, e não o seu erro, na distribuição do atributo principal do conjunto de dados: o tipo de parto.

Para tal, foi construída a tabela de contingência (*cross table*) com as contagens de casos para cada valor das variáveis e suas participações nos tipos de parto, Tabela 6. Os tipos de parto foram resumidos em parto cesariana, parto eutócico (normal) e parto auxiliado (instrumentado) para esta conferência dos valores. Foram abstraídos os tipos de instrumentação (partos auxiliados com fórceps e/ou ventosa). Na análise do total de registros dos hospitais foi calculada a significância de cada variável relativa ao tipo de parto com os testes $\tilde{\chi}^2$ (*Chi squared* ou qui quadrado) de Pearson. O teste de Pearson, ou $\tilde{\chi}^2$, foi utilizado em amostras com frequências superiores a 20% da frequência esperada em observações por caso em tabelas de contingência. Isso é assim definido porque, em pequenas amostras, a aproximação da distribuição do χ é afetada e não considerada satisfatória. Para os demais casos foi utilizado o teste exato de Fisher (CHRISTENSEN, 2005). Estes casos estão identificados na tabela. Para os testes de significância, foram desconsiderados os valores faltantes e casos absolutamente raros do tipo de parto "grande extração pélvica". O total de registros dos conjuntos de dados combinados é de 25.945. Nas variáveis que não possuem dados omitidos pela falta de preenchimento, o total de registros preenchidos da variável na Tabela 6 está omitido por ser igual ao total.

Tabela 6 – Frequência, $p-value$ e contribuição ao tipo de parto nas variáveis no total da amostra de hospitais portugueses e significância entre estas variáveis

	Tipo de parto			**Total**	
	Cesariana	**Parto auxiliado**	**Parto eutócico**	25.945 (100%)	**p-value**
Grupo de Robson	7.574 (29%)	4.701 (18%)	13.670 (53%)	-	
1	1.121 (4,3%)	1.958 (7,5%)	3.438 (13%)	6.517 (25%)	
2	1.527 (5,9%)	1.034 (4,0%)	1.462 (5,6%)	4.023 (16%)	
3	211 (0,8%)	433 (1,7%)	4.699 (18%)	5.343 (21%)	
4	362 (1,4%)	215 (0,8%)	1.530 (5,9%)	2.107 (8,1%)	
5	1.722 (6,6%)	537 (2,1%)	850 (3,3%)	3.109 (12%)	
6	861 (3,3%)	0 (0%)	10 (<0,1%)	871 (3,4%)	**<0,001**
7	522 (2,0%)	0 (0%)	38 (0,1%)	560 (2,2%)	
8	357 (1,4%)	49 (0,2%)	110 (0,4%)	516 (2,0%)	
9	29 (0,1%)	0 (0%)	0 (0%)	29 (0,1%)	
10	862 (3,3%)	475 (1,8%)	1.533 (5,9%)	2.870 (11%)	
Número de RN	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
1	7,217 (28%)	4,652 (18%)	13,560 (52%)	25,429 (98%)	
2	340 (1.3%)	49 (0.2%)	109 (0.4%)	498 (1.9%)	**<0.001** **
3	17 (<0.1%)	0 (0%)	1 (<0.1%)	18 (<0.1%)	
Semanas de gestação	7,568 (29%)	4,697 (18%)	13,638 (53%)	-	
$\leq Q_1$	2,157 (8.3%)	858 (3.3%)	3,555 (14%)	6,570 (25%)	
$> Q_1 \& \leq Q_2$	2,016 (7.8%)	1,061 (4.1%)	3,498 (14%)	6,575 (25%)	**<0.001**
$> Q_2 \& \leq Q_3$	1,674 (6.5%)	1,453 (5.6%)	3,884 (15%)	7,011 (27%)	
$> Q_3$	1,721 (6.6%)	1,325 (5.1%)	2,701 (10%)	5,747 (22%)	
Semanas de gestação (admissão)	7,515 (29%)	4,668 (18%)	13,516 (53%)	25,699 (100%)	
$\leq Q_1$	2,268 (8.8%)	921 (3.6%)	3,691 (14%)	6,880 (27%)	
$> Q_1 \& \leq Q_2$	2,040 (7.9%)	1,110 (4.3%)	3,514 (14%)	6,664 (26%)	
$> Q_2 \& \leq Q_3$	1,443 (5.6%)	1,220 (4.7%)	3,303 (13%)	5,966 (23%)	**<0.001**
$> Q_3$	1,764 (6.9%)	1,417 (5.5%)	3,008 (12%)	6,189 (24%)	
Apresentação (no parto)	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
Cefálica	6,006 (23%)	4,701 (18%)	13,609 (52%)	24,316 (94%)	
Desconhecida	19 (<0.1%)	0 (0%)	0 (0%)	19 (<0.1%)	**<0.001** **
Pélvica	1,507 (5.8%)	0 (0%)	61 (0.2%)	1,568 (6.0%)	
Transversa	42 (0.2%)	0 (0%)	0 (0%)	42 (0.2%)	
Apresentação (admissão)	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
Cefálica	821 (3.2%)	798 (3.1%)	2,600 (10%)	4,219 (16%)	
Desconhecida	289 (1.1%)	124 (0.5%)	318 (1.2%)	731 (2.8%)	
Pélvica	177 (0.7%)	1 (<0.1%)	10 (<0.1%)	188 (0.7%)	**<0.001** **
Transversa	11 (<0.1%)	4 (<0.1%)	10 (<0.1%)	25 (<0.1%)	
Não preenchido	6,276 (24%)	3,774 (15%)	10,732 (41%)	20,782 (80%)	
Trabalho de parto (no parto)	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
Ausência de trabalho de parto	2,323 (9.0%)	1 (<0.1%)	5 (<0.1%)	2,329 (9.0%)	
Espontâneo	2,939 (11%)	2,961 (11%)	9,636 (37%)	15,536 (60%)	**<0.001**
Indução	2,312 (8.9%)	1,739 (6.7%)	4,029 (16%)	8,080 (31%)	
TP espontâneo (admissão)	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
Não preenchido	6,222 (24%)	3,046 (12%)	8,335 (32%)	17,603 (68%)	**<0.001**
SIM	1,352 (5.2%)	1,655 (6.4%)	5,335 (21%)	8,342 (32%)	
TP induzido (admissão)	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
Não preenchido	6,921 (27%)	4,232 (16%)	12,548 (48%)	23,701 (91%)	**<0.001**
Sim	653 (2.5%)	469 (1.8%)	1,122 (4.3%)	2,244 (8.6%)	
Nº de filhos	6,157 (26%)	4,365 (19%)	12,957 (55%)	23,479 (100%)	
0	2,914 (12%)	3,045 (13%)	5,042 (21%)	11,001 (47%)	
1	2,540 (11%)	1,166 (5.0%)	6,093 (26%)	9,799 (42%)	
2	552 (2.4%)	118 (0.5%)	1,377 (5.9%)	2,047 (8.7%)	
3	117 (0.5%)	29 (0.1%)	321 (1.4%)	467 (2.0%)	
4	22 (<0.1%)	7 (<0.1%)	85 (0.4%)	114 (0.5%)	**<0.001** **
5	5 (<0.1%)	0 (0%)	28 (0.1%)	33 (0.1%)	
6	2 (<0.1%)	0 (0%)	8 (<0.1%)	10 (<0.1%)	
7	4 (<0.1%)	0 (0%)	2 (<0.1%)	6 (<0.1%)	
8	1 (<0.1%)	0 (0%)	0 (0%)	1 (<0.1%)	
9	0 (0%)	0 (0%)	1 (<0.1%)	1 (<0.1%)	

Tabela 7 – (continuação tabela 6)

	Tipo de parto			Total	
	Cesariana	**Parto auxiliado**	**Parto eutócico**	25.945 (100%)	**p-value**
Nº de gestações	7,573 (29%)	4,700 (18%)	13,669 (53%)	25,942 (100%)	
0	2 (<0.1%)	0 (0%)	3 (<0.1%)	5 (<0.1%)	
1	3,492 (13%)	2,855 (11%)	4,823 (19%)	11,170 (43%)	
2	2,460 (9.5%)	1,303 (5.0%)	5,329 (21%)	9,092 (35%)	
3	1,062 (4.1%)	374 (1.4%)	2,278 (8.8%)	3,714 (14%)	
4	344 (1.3%)	105 (0.4%)	796 (3.1%)	1,245 (4.8%)	
5	144 (0.6%)	46 (0.2%)	271 (1.0%)	461 (1.8%)	
6	39 (0.2%)	11 (<0.1%)	106 (0.4%)	156 (0.6%)	**<0.001 ****
7	14 (<0.1%)	2 (<0.1%)	38 (0.1%)	54 (0.2%)	
8	12 (<0.1%)	2 (<0.1%)	16 (<0.1%)	30 (0.1%)	
9	4 (<0.1%)	1 (<0.1%)	5 (<0.1%)	10 (<0.1%)	
10	0 (0%)	1 (<0.1%)	1 (<0.1%)	2 (<0.1%)	
11	0 (0%)	0 (0%)	1 (<0.1%)	1 (<0.1%)	
16	0 (0%)	0 (0%)	1 (<0.1%)	1 (<0.1%)	
17	0 (0%)	0 (0%)	1 (<0.1%)	1 (<0.1%)	
Cesarianas anteriores	5,530 (25%)	4,220 (19%)	12,169 (56%)	21,919 (100%)	
0	3,406 (16%)	3,654 (17%)	11,236 (51%)	18,296 (83%)	
1	1,781 (8.1%)	564 (2.6%)	931 (4.2%)	3,276 (15%)	
2	306 (1.4%)	1 (<0.1%)	2 (<0.1%)	309 (1.4%)	**<0.001 ****
3	33 (0.2%)	1 (<0.1%)	0 (0%)	34 (0.2%)	
4	4 (<0.1%)	0 (0%)	0 (0%)	4 (<0.1%)	
Gravidez vigiada	7,573 (29%)	4,698 (18%)	13,663 (53%)	25,934 (100%)	
Mal vigiada	1 (<0.1%)	0 (0%)	0 (0%)	1 (<0.1%)	
Não	232 (0.9%)	158 (0.6%)	624 (2.4%)	1,014 (3.9%)	**<0.001**
Não sei	9 (<0.1%)	6 (<0.1%)	29 (0.1%)	44 (0.2%)	
Sim	7,331 (28%)	4,534 (17%)	13,010 (50%)	24,875 (96%)	
Consultas pré natal	5,766 (30%)	3,612 (19%)	9,776 (51%)	19,154 (100%)	
$\leq Q_1$	1,390 (7.3%)	1,028 (5.4%)	2,715 (14%)	5,133 (27%)	
$> Q_1 \& \leq Q_2$	1,748 (9.1%)	1,115 (5.8%)	3,025 (16%)	5,888 (31%)	**<0.001**
$> Q_2 \leq Q_3$	1,510 (7.9%)	903 (4.7%)	2,484 (13%)	4,897 (26%)	
$> Q_3$	1,118 (5.8%)	566 (3.0%)	1,552 (8.1%)	3,236 (17%)	
Idade materna	7,574 (29%)	4,701 (18%)	13,670 (53%)	-	
$>\mu + 1*\sigma$	1,587 (6.1%)	593 (2.3%)	2,120 (8.2%)	4,300 (17%)	
$> \mu - 1*\sigma < x \leq \mu$	2,425 (9.3%)	1,815 (7.0%)	4,731 (18%)	8,971 (35%)	**<0.001**
$\mu < x \leq \mu + 1*\sigma$	2,596 (10%)	1,354 (5.2%)	4,360 (17%)	8,310 (32%)	
$x \leq \mu - 1*\sigma$	966 (3.7%)	939 (3.6%)	2,459 (9.5%)	4,364 (17%)	
Peso inicial	5,398 (28%)	3,671 (19%)	10,397 (53%)	19,466 (100%)	
$\leq Q_1$	1,302 (6.7%)	1,073 (5.5%)	2,705 (14%)	5,080 (26%)	
$> Q_1 \& \leq Q_2$	1,261 (6.5%)	958 (4.9%)	2,616 (13%)	4,835 (25%)	**<0.001**
$> Q_2 \leq Q_3$	1,294 (6.6%)	884 (4.5%)	2,522 (13%)	4,700 (24%)	
$> Q_3$	1,541 (7.9%)	756 (3.9%)	2,554 (13%)	4,851 (25%)	
IMC (graus)	7,574 (29%)	4,701 (18%)	13,670 (53%)		
Abaixo do peso	128 (0.5%)	110 (0.4%)	380 (1.5%)	618 (2.4%)	
Obesidade grau I	639 (2.5%)	282 (1.1%)	967 (3.7%)	1,888 (7.3%)	
Obesidade grau II	248 (1.0%)	104 (0.4%)	314 (1.2%)	666 (2.6%)	**<0.001**
Obesidade grau III	2,410 (9.3%)	1,145 (4.4%)	3,659 (14%)	7,214 (28%)	
Peso normal	2,705 (10%)	2,227 (8.6%)	5,754 (22%)	10,686 (41%)	
Sobrepeso	1,444 (5.6%)	833 (3.2%)	2,596 (10%)	4,873 (19%)	
Cigarro	7,574 (29%)	4,701 (18%)	13,670 (53%)		
Não	6,726 (26%)	4,189 (16%)	11,939 (46%)	22,854 (88%)	**<0.001**
Sim	848 (3.3%)	512 (2.0%)	1,731 (6.7%)	3,091 (12%)	
Álcool	7,574 (29%)	4,701 (18%)	13,670 (53%)		
Não	7,423 (29%)	4,627 (18%)	13,414 (52%)	25,464 (98%)	0.2
Sim	151 (0.6%)	74 (0.3%)	256 (1.0%)	481 (1.9%)	
Estupefacientes	7,574 (29%)	4,701 (18%)	13,670 (53%)		
Não	7,558 (29%)	4,691 (18%)	13,650 (53%)	25,899 (100%)	0.5
Sim	16 (<0.1%)	10 (<0.1%)	20 (<0.1%)	46 (0.2%)	
Tipo sanguíneo	7,551 (29%)	4,691 (18%)	13,601 (53%)	25,843 (100%)	
A	3,748 (15%)	2,257 (8.7%)	6,484 (25%)	12,489 (48%)	
AB	275 (1.1%)	185 (0.7%)	520 (2.0%)	980 (3.8%)	0.10
B	599 (2.3%)	391 (1.5%)	1,058 (4.1%)	2,048 (7.9%)	
O	2,929 (11%)	1,858 (7.2%)	5,539 (21%)	10,326 (40%)	

Tabela 8 – (continuação tabela 7)

	Tipo de parto			**Total**	
	Cesariana	**Parto auxiliado**	**Parto eutócico**	25.945 (100%)	**p-value**
Tipo de gravidez	6,478 (29%)	3,933 (18%)	11,633 (53%)	22,044 (100%)	
Espontânea	6,073 (28%)	3,777 (17%)	11,332 (51%)	21,182 (96%)	
Esimulação	36 (0.2%)	18 (<0.1%)	41 (0.2%)	95 (0.4%)	
FIV	194 (0.9%)	72 (0.3%)	137 (0.6%)	403 (1.8%)	**<0.001**
ICSI	85 (0.4%)	27 (0.1%)	63 (0.3%)	175 (0.8%)	
Inseminação	39 (0.2%)	18 (<0.1%)	32 (0.1%)	89 (0.4%)	
TEC	51 (0.2%)	21 (<0.1%)	28 (0.1%)	100 (0.5%)	
Diabetes gestacional	7,574 (29%)	4,701 (18%)	13,670 (53%)		
Não preenchido	6,746 (26%)	4,274 (16%)	12,415 (48%)	23,435 (90%)	**<0.001**
Sim	828 (3.2%)	427 (1.6%)	1,255 (4.8%)	2,510 (9.7%)	
Doenças cardíacas	7,574 (29%)	4,701 (18%)	13,670 (53%)		
Não preenchido	6,880 (27%)	4,473 (17%)	12,997 (50%)	24,350 (94%)	**<0.001**
Sim	694 (2.7%)	228 (0.9%)	673 (2.6%)	1,595 (6.1%)	

Quanto maior a possibilidade de valores em uma variável (para as redes Bayesianas) menor será a importância e a confiança da descrição que a rede estará fazendo sobre a variável. A partir da análise da tabela de contingência das variáveis do conjunto de dados, algumas reclassificações se fazem necessárias para diminuir a magnitude dos valores possíveis e agregar conhecimento para o estudo. Os grupos de Robson, segundo a interpretação de Robson, podem ser classificados em cesarianas muito evitáveis (grupos 1, 2, 3 e 4), evitáveis (5, 6 e 9) e pouco evitáveis (7, 8 e 10) (ROBSON, 2001). A variável sobre o número de gestações, filhos e cesarianas anteriores também sofreu mudanças segundo as tabelas a seguir. Os valores apresentados por quartis e médias nas variáveis quantitativas e contínuas também se mantiveram desta forma. A tabela de contingência, ou tabela cruzada, é apresentada na Tabela 9. As variáveis reclassificadas e/ou com significância ($p - value < 0{,}001$), foram incluídas no estudo do preenchimento de dados por técnicas computacionais e para construção dos modelos.

Sobre essas variáveis, são identificados os mecanismos em que os *missings* possam ter sido gerados (MCAR, MNAR ou MAR). E, pelos mecanismos, foi proposto a preenchimento pelo MICE. A comparação sobre os preenchimentos e os dados reais é realizada quanto à distribuição dos dados pelos grupos de Robson.

Tabela 9 – Contribuição do tipo de parto nas classes de cesarianas no total da amostra de hospitais portugueses e significância entre estas variáveis

CLASS_CESA	PARTO			Total	p-value*
	Cesariana	Parto auxiliado	Parto eutócico		
CS evitável	2,612 (10%)	537 (2.1%)	860 (3.3%)	4,009 (15%)	<0.001
CS muito evitável	3,221 (12%)	3,640 (14%)	11,129 (43%)	17,990 (69%)	
CS pouco evitável	1,741 (6.7%)	524 (2.0%)	1,681 (6.5%)	3,946 (15%)	
Total	7,574 (29%)	4,701 (18%)	13,670 (53%)	25,945 (100%)	

* Pearson's Chi-squared test

Tabela 10 – Contribuição do tipo de parto no número de gestações no total da amostra de hospitais portugueses e significância entre estas variáveis

Nº gestações	PARTO			Total	p-value*
	Cesariana	Parto auxiliado	Parto eutócico		
1	3,494 (13%)	2,855 (11%)	4,826 (19%)	11,175 (43%)	<0.001
2	2,460 (9.5%)	1,303 (5.0%)	5,329 (21%)	9,092 (35%)	
3	1,062 (4.1%)	374 (1.4%)	2,278 (8.8%)	3,714 (14%)	
4 ou mais	557 (2.1%)	168 (0.6%)	1,236 (4.8%)	1,961 (7.6%)	
Total	7,573 (29%)	4,700 (18%)	13,669 (53%)	25,942 (100%)	

* Pearson's Chi-squared test

Tabela 11 – Contribuição do tipo de parto no número de filhos no total da amostra de hospitais portugueses e significância entre estas variáveis

Nº de filhos	PARTO			Total	p-value*
	Cesariana	Parto auxiliado	Parto eutócico		
0	2,914 (12%)	3,045 (13%)	5,042 (21%)	11,001 (47%)	<0.001
1	2,540 (11%)	1,166 (5.0%)	6,093 (26%)	9,799 (42%)	
2 ou mais	703 (3.0%)	154 (0.7%)	1,822 (7.8%)	2,679 (11%)	
Total	6,157 (26%)	4,365 (19%)	12,957 (55%)	23,479 (100%)	

* Pearson's Chi-squared test

Tabela 12 – Contribuição do tipo de parto no número de cesarianas anteriores no total da amostra de hospitais portugueses e significância entre estas variáveis

Cesarianas anteriores	PARTO			Total	p-value*
	Cesariana	Parto auxiliado	Parto eutócico		
0	3,406 (16%)	3,654 (17%)	11,236 (51%)	18,296 (83%)	<0.001
1	1,781 (8.1%)	564 (2.6%)	931 (4.2%)	3,276 (15%)	
2 ou mais	343 (1.6%)	2 (<0.1%)	2 (<0.1%)	347 (1.6%)	
Total	5,530 (25%)	4,220 (19%)	12,169 (56%)	21,919 (100%)	

* Pearson's Chi-squared test

5.7 Modelo representativo dos dados obstétricos

Após definidas as variáveis com significância aos grupos de tipo de parto, variáveis com mais de três tipos de valores foram reclassificadas em categorias para a construção da rede Bayesiana. São elas:

"N_CONSULTAS_CATEGORIAS": quantidade de consultas pré-natal,

"IDADE_MATERNA_CATEGORIAS": idade materna dividida em categorias segundo média e desvio padrão,

"IMC_CATEGORIAS": categorias de IMC segundo graus de obesidade,

"A_PARA_CATEGORIAS": quantidade de filhos nascidos,

"A_CESA_CATEGORIAS": quantidade de cesarianas realizadas,

"A_GESTA_CATEGORIAS": quantidade de gestações, incluindo abortos e a atual,

"SEMANAS_CATEGORIAS": categorias das semanas de gestação segundo quartis,

"NUM_RN": quantidade de fetos na gestação atual

"APRESENTACAO_NO_PARTO": apresentação do feto na hora do parto e

"TRAB_PARTO_NO_PARTO": trabalho de parto na hora do parto.

Para cada hospital e seleção de variáveis que apresentaram significância ao tipo de parto referentes à admissão ao parto (variáveis terminadas em "ADM"nos modelos), ao momento do parto, gravidez e características da grávida, foi criada uma rede Bayesiana. Para criar o modelo, apenas os registros com dados completos são selecionados. Logo, com os dados não preenchidos, a quantidade de registros diminui. Para cada hospital, o volume do conjunto de dados e seleção dos registros completos é descrito na Tabela 13.

Tabela 13 – Total de registros por hospital e total de registros completos

	Total	Registros completos
H1	11.705	4.219 (36,0%)
H2	7.743	3.922 (50,7%)
H3	6.572	3.110 (47,3%)
Geral	26.020	11.251 (43,2%)

As redes Bayesianas geradas com estes dados são apresentadas nas Figuras 13, 14 e 15. Nota-se que os arcos não tiveram o mesmo comportamento. Estas dissemelhanças estruturais são utilizadas para calcular a distância Hamming entre os modelos. Pela falta de informação na seleção dos registros completos, o hospital H3 não obteve informação suficiente para mapear os 23 nodos, removendo os nodos SEMANAS_ADM, APRESENTACAO_ADM, TP_ESPONTANEO_ADM, TP_INDUZIDO_ADM, N_CONSULTAS, PESO_INICIAL.

A partir dos modelos, foram recriados dados com a mesma quantidade de registros dos conjuntos reais. Para fins de comparação, foram extraídos moda e participação de cada valor da variável no total de registros das variáveis categóricas (Tabela 14) e na variável quantitativa, nomeadamente na variável referente a quantidade de bebês na gestação (NUM_RN) foram extraídos a média, mínimo e máximo. A comparação das medidas na variável quantitativa das variáveis do conjunto de dados reais e do produzido é apresentada na Tabela 15 e a tabela de contingências com a moda das variáveis categóricas por conjunto de dados reais, dados completos e dados recriados de cada hospital na Tabela 14.

Tabela 14 – Moda das variáveis selecionadas nos conjuntos de dados reais, seleção dos registros completos e dados recriados

Hospitais	H1			H2			H3		
	moda			moda			moda		
Variáveis	dados reais	completos	dados recriados	dados reais	completos	recriados	dados reais	completos	recriados
SEMANAS_ADM	$>Q_1 \& \leq Q_2$	$>Q_3$	$>Q_3$	$>Q_1 \& \leq Q_2$	$>Q_2 \leq Q_3$	$>Q_2 \leq Q_3$	$\leq Q_1$	$>Q_2 \leq Q_3$	-
Apresentação (admissão)	não preenchido			não preenchido			não preenchido		-
TP espontâneo (admissão)	não preenchido			não preenchido			não preenchido		-
TP induzido (admissão)	não preenchido			não preenchido			não preenchido		-
N_CONSULTAS	*NA*	$\leq Q_1$	$\leq Q_1$	$\leq Q_1$			$>Q_1 \& \leq Q_2$		-
PESO_INICIAL	*NA*	$>Q_3$	$>Q_1 \& \leq Q_2$	$\leq Q_1$	$>Q_3$	$>Q_3$	$>Q_1 \& \leq Q_2$	$\leq Q_1$	-
VIGIADA	S			S			S		
IDADE_MATERNA	$\mu < x \leq \mu + 1*\sigma$			$\mu < x \leq \mu + 1*\sigma$			$> \mu - 1*\sigma < x \leq \mu$	$\mu < x \leq \mu + 1*\sigma$	$\mu < x \leq \mu + 1*\sigma$
Diabetes gestacional	não preenchido			não preenchido			não preenchido		
Doenças cardíacas	não preenchido			não preenchido			não preenchido		
IMC	*NA*	Peso normal	Peso normal	Peso normal			Peso normal		
Cigarros	NAO			NAO			NAO		
Tipo gravidez	ESPONTANEA			ESPONTANEA			ESPONTANEA		
CS anteriores	0			0			0		
Gestações	1			1			1		
Semanas de gestação	$>Q_1 \& \leq Q_2$	$>Q_2 \leq Q_3$	$>Q_3$	$>Q_1 \& \leq Q_2$	$>Q_2 \leq Q_3$	$>Q_2 \leq Q_3$	$>Q_1 \& \leq Q_2$	$>Q_2 \leq Q_3$	
Nº de RN	1			1			1		
Apresentação	Cefálica			Cefálica			Cefálica		
TP	Espontâneo			Espontâneo			Espontâneo		
Tipo de parto	Parto eutócico			Parto eutócico			Parto eutócico		
Nº filhos	0			0			0		
Grupo de Robson	1			1			1		
Classe de CS	CS muito evitável			CS muito evitável			CS muito evitável		

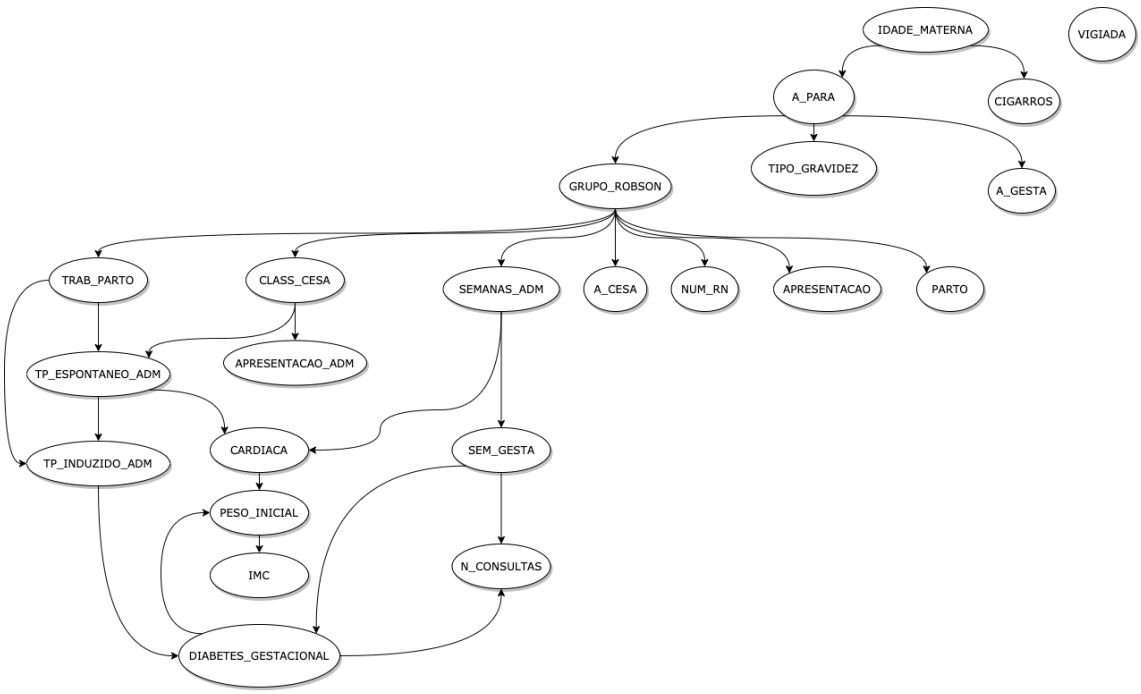


Figura 13 – Rede Bayesiana produzida por meio dos dados do hospital H1

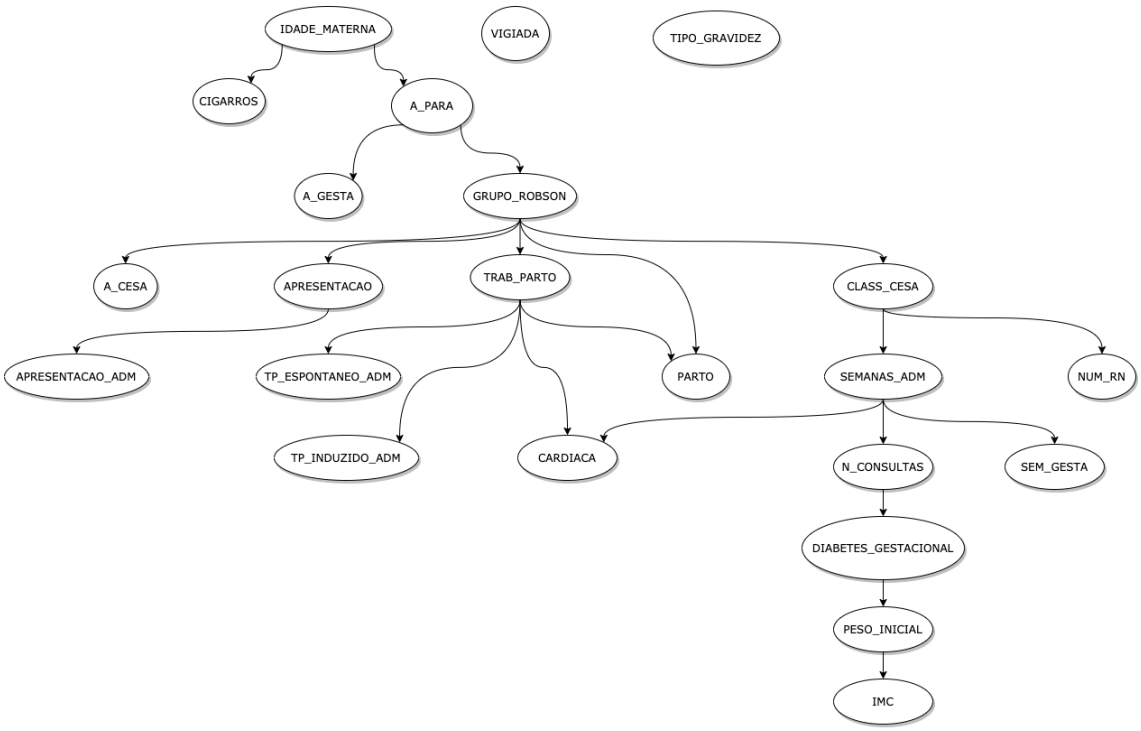


Figura 14 – Rede Bayesiana produzida por meio dos dados do hospital H2

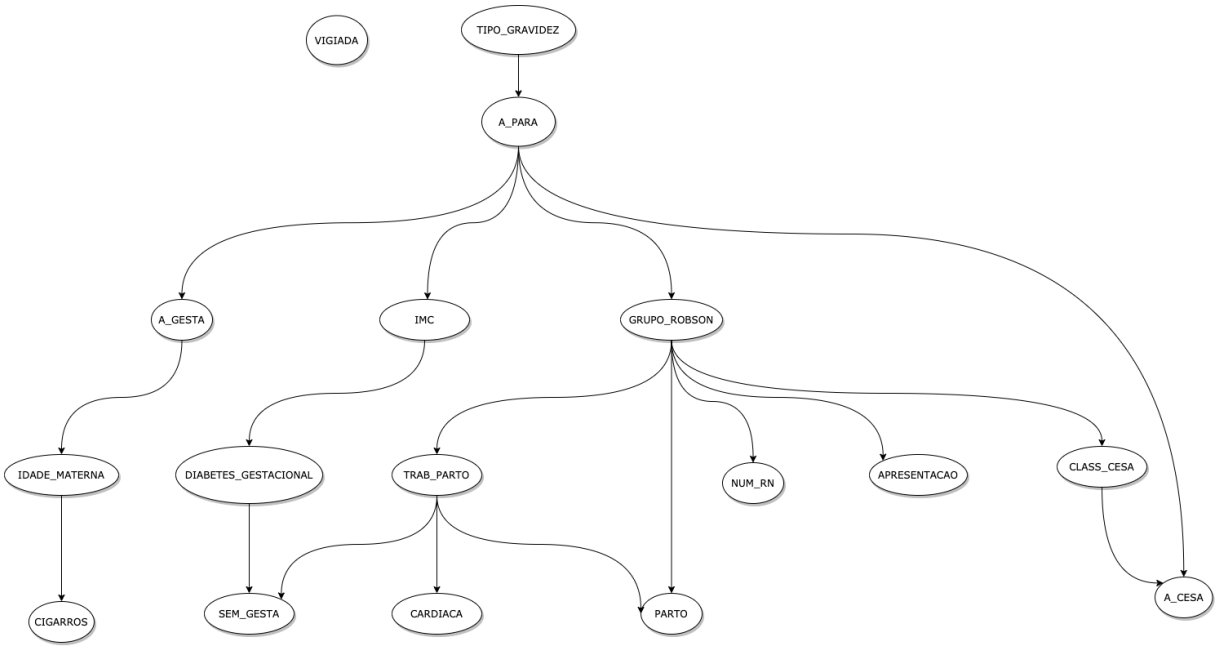


Figura 15 – Rede Bayesiana produzida por meio dos dados do hospital H3

Tabela 15 – Características dos dados referentes a quantidade de bebês na gestação (variável *NUM_RN*) nos conjuntos de dados dos hospitais (média [min, max])

Hospitais	**H1**	**H2**	**H3**
Dados reais	1.02 [1 , 3]	1.01 [1, 3]	1.02 [1 , 3]
Dados reais completos	1.01 [1, 3]	1.00 [1, 3]	1.00 [1, 2]
Dados recriados	1.01 [1, 3]	1.00 [1, 3]	1.00 [1, 2]

5.8 Detecção de anomalia

Como é impossível detectar uma anomalia a partir de três observações ou entidades, nesse caso três hospitais, foram feitas subamostragens de cada hospital com o mesmo volume de dados dos dados originais. Afim de iniciar uma comparação do total de hospitais, sabe-se que havia 80 hospitais na região Norte de Portugal em 2019 (Fontes/Entidades: INE | DGS/MS, PORDATA) e, destes, 33 são hospitais públicos (PORTUGAL, 2020). Considerando que a quantidade de hospitais não varia significativamente nos anos, os 3 hospitais do caso de estudo representam apenas 3,75% do total de hospitais do Norte de Portugal, 9,09% dos hospitais públicos da mesma região. Para simular esta quantidade de hospitais públicos foram gerados 11 conjuntos de dados por modelo (total das 33 entidades a serem comparadas) a partir dos três modelos de hospitais criados. Com as 33 entidades foram construídos novos 33 modelos a serem comparados.

Além destes, foi incluída à análise os modelos gerados com a seleção dos dados reais completos dos hospitais H2 e H1. A comparação foi incluída na mesma proporção, somando outros 11 modelos gerados a partir dos dados criados pelos modelos com dados completos, somando 22 hospitais à comparação. O modelo gerado pelos dados reais do hospital H3 não foi agregado a esta análise por não conter todos os nodos representados e, por esta razão, a contagem de arcos e direções dos arcos dissemelhantes não podem ser contabilizadas.

A dissimmilaridade entre os modelos criados foi calculada através da distância Hamming de modelo a modelo. Com essas distâncias foi construída a matriz de distâncias ou matriz de dissimilaridades entre os modelos. Na matriz de distâncias encontram-se todas as distâncias de cada hospital para todos outros e, a partir da manipulação dela, são executados os algoritmos de detecção de anomalia.

CAPÍTULO

6

RESULTADOS

6.1 Preenchimento de dados

De acordo com a identificação dos mecanismos, as variáveis do conjunto de dados correspondem ao mecanismo MAR, *missing at random* ou valores faltantes completamente aleatórios. Isso indica que a razão dos valores faltantes é exclusivamente dos dados observados, e não razão não observada. Por esta razão, técnicas de preenchimento que não relacionam variáveis do conjunto podem resultar num bom preenchimento. Serão aplicadas as técnicas de preenchimento por média e moda, de acordo com o tipo de variável quantitativa/categórica e a técnica *Multiple Imputation by Chained Equations* (MICE) (ROYSTON; WHITE *et al.*, 2011).

Os dados preenchidos e os dados reais são comparados quanto a distribuição nos grupos de Robson na Tabela 16. Nessa tabela, as colunas $input1$ referem-se aos dados preenchidos com média nas variáveis quantitativas e moda nas variáveis qualitativas e $input2$ referem-se ao preenchimento pela abordagem MICE. Na comparação, não houve diferença considerada significativa na distribuição de partos nos grupos de Robson entre os dados reais, os dados preenchidos pela média e moda e os preenchidos pela técnica MICE para cada variável do conjunto de dados. Ambas resultaram em dados de características muito próximas aos dados reais, porém, por comparação direta na Tabela 16 , a técnica que obteve a maior similaridade foi a $input2$, com o algoritmo MICE.

Tabela 16 – Tabela de contingência com distribuição da frequência das participações em porcentagem de registros nos grupos de Robson (%) no total de dados reais sem preenchimento, dos dados preenchidos pela média e moda (input1) e preenchidos com a abordagem MICE (input2)

	H1			H2			H3		
GR	reais	input1	input2	reais	input1	input2	reais	input1	input2
1	27,1	27,2	27,1	22,3	21,3	22,3	24,5	24,1	24,5
2	14,1	14,3	14,1	18,0	18,3	18,0	14,8	14,7	14,8
3	20,6	20,7	20,6	20,8	21,2	20,8	20,0	19,6	20,0
4	5,8	5,6	5,8	10,1	10,4	10,1	9,8	10,4	9,8
5	13,2	13,1	13,2	12,1	12,4	12,1	9,5	8,8	9,5
6	3,8	4,1	3,8	2,9	2,6	2,9	3,1	3,2	3,1
7	2,3	2,5	2,3	1,9	1,9	1,9	2,2	2,5	2,2
8	2,1	2,0	2,0	1,5	1,6	1,5	2,4	2,5	2,4
9	0,1	0,1	0,1	0,2	0,2	0,2	0,0	0,1	0,0
10	10,5	10,1	10,5	10,2	10,2	10,2	13,0	13,8	13,0
U	0,4	0,4	0,4	0,0	0,0	0,0	0,3	0,4	0,3

6.2 Modelos gerados

Para criar a estrutura e as probabilidades das redes Bayesianas de cada hospital é requerida a completude dos dados. Por isso, foram selecionados os registros com os valores completos das variáveis selecionadas e criados modelos a partir deles. Os modelos gerados pelos dados reais de cada hospital são as RB apresentadas no caso de estudo nas Figuras 13, 14 e 15. Nota-se no modelo representativo de H3 (Figura 15) que, na criação automática do modelo a partir dos dados reais, não foram adicionados os nodos das variáveis APRESENTACAO_ADM, TP_ESPONTANEO_ADM, TP_INDUZIDO_ADM, SEMANAS_ADM, N_CONSULTAS e PESO_INICIAL.

A partir desses modelos foram recriados os dados. Foram comparadas as distribuições dos partos nos grupos de Robson nos dados reais, na seleção dos dados completos e nos dados recriados pelos modelos na Tabela 17. Ao comparar a distribuição dos partos pelos grupos nos diferentes conjuntos é possível observar grande diferença dos conjuntos reais para a seleção dos dados completos, ressaltadas na tabela. Esta diferença é refletida quando gerado um modelo a partir da seleção dados completos e recriados dados deste modelo.

Tabela 17 – Tabela de contingência com distribuição da frequência das participações em porcentagem de registros pelos grupos de Robson (%) no total de dados originais (N) e total de registros completos (n). Os dados recriados possuem o tamanho dos dados originais.

	H1			**H2**			**H3**		
	N=11.705 , n=4.219 (36,0%)			N=7.743 , n=3.922 (50,7%)			N=6.572 , n= 3.110 (47,3%)		
GR	real	real completo	recriado	real	real completo	recriado	real	real completo	recriado
1	**27,1**	31,8	31,3	**22,3**	28,4	28,7	**24,5**	29,7	30,6
2	**14,1**	17,8	17,9	18,0	19,7	19,9	**14,8**	20,0	19,7
3	**20,6**	23,5	23,6	**20,8**	25,0	24,6	**20,0**	24,0	23,9
4	5,8	7,3	7,3	10,1	10,5	10,5	**9,8**	13,0	11,9
5	13,2	14,9	15,1	12,1	13,7	13,5	9,5	11,2	11,7
6	**3,8**	0,8	0,9	**2,9**	0,3	0,3	**3,1**	0,0	0,0
7	2,3	1,1	1,2	1,9	0,9	1,0	2,2	0,7	0,6
8	2,1	1,3	1,1	1,5	0,5	0,5	2,4	0,5	0,6
9	0,1	0,0	0,0	0,2	0,1	0,1	0,0	0,0	0,0
10	**10,5**	1,5	1,5	**10,2**	0,9	0,9	**13,0**	1,0	1,0
U	0,4	0,0	0	0,0	0	0,0	0,3	0,0	0,0

Foram também criados modelos com os dados preenchidos pela média e moda e preenchidos por MICE. A estrutura dos modelos gerados pelos conjuntos de dados inserindo os valores faltantes pelas duas técnicas de preenchimento foram iguais e são apresentados nas Figuras 16, 17 e 18. Ao preencher os dados, os nodos omitidos ao selecionar os dados completos voltam a ser representados nos modelos gerados. Como concluído anteriormente, os dados que geraram estes modelos produziram uma distribuição semelhante a dos dados reais, independente do tipo de preenchimento (discutidos na sessão 6.1). Por este motivo, os dados preenchidos por média e moda foram utilizados para a comparação e detecção de anomalia.

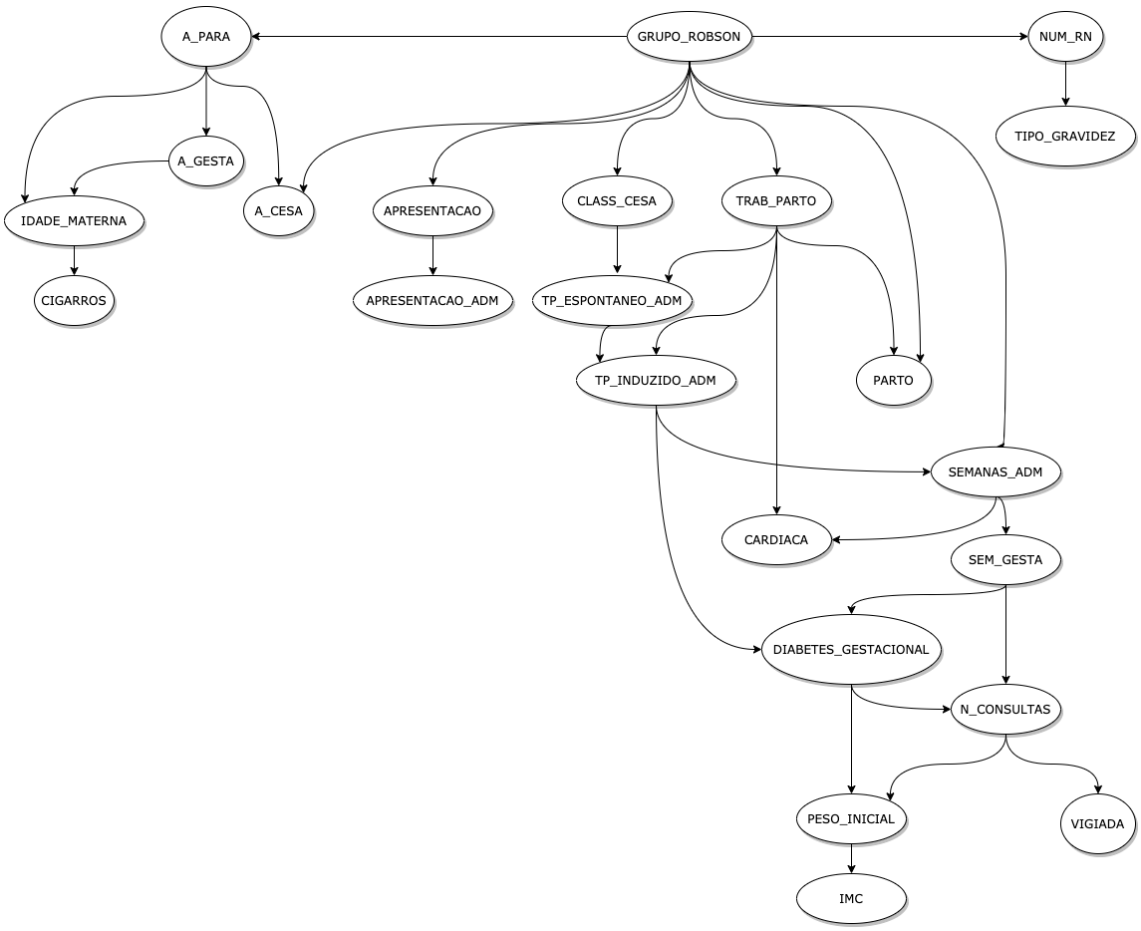


Figura 16 – Modelo gerado pelos dados do hospital H1 com dados preenchidos com média e moda das variáveis

Figura 17 – Modelo gerado pelos dados do hospital H2 com dados preenchidos com média e moda das variáveis

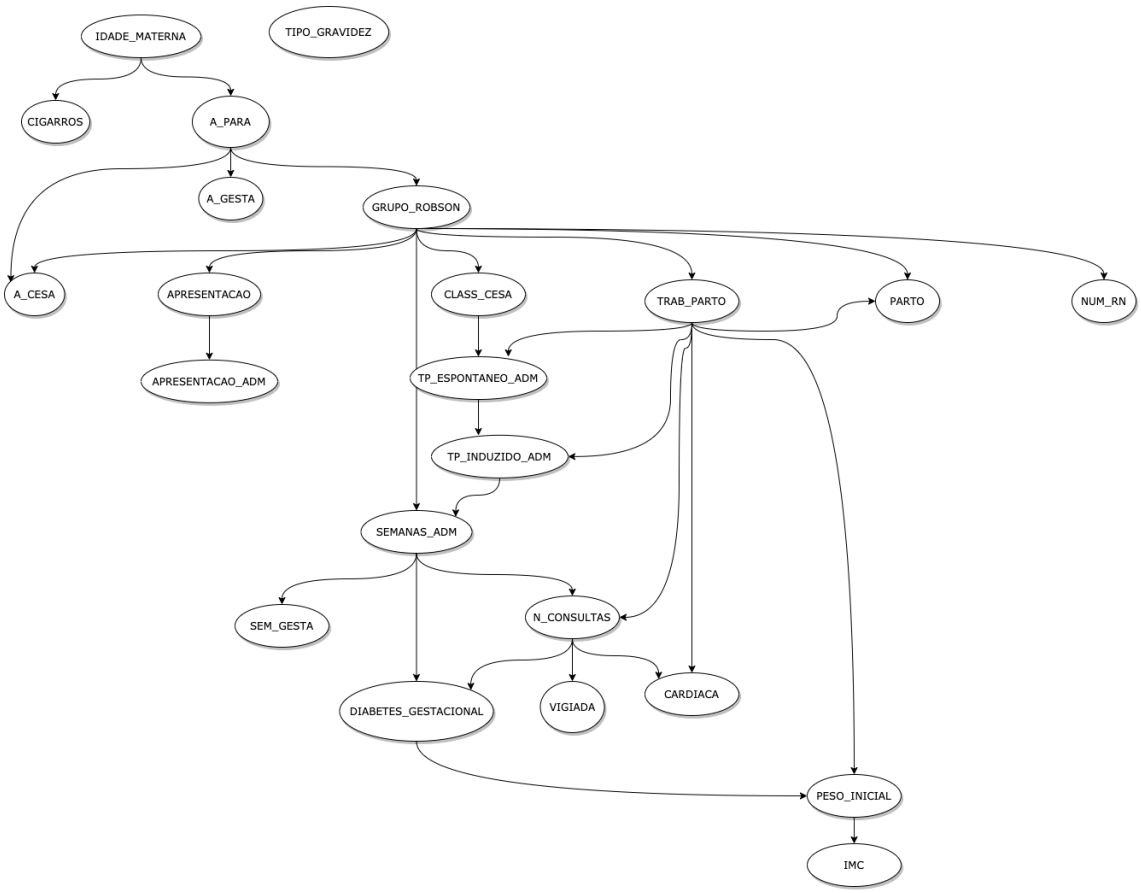
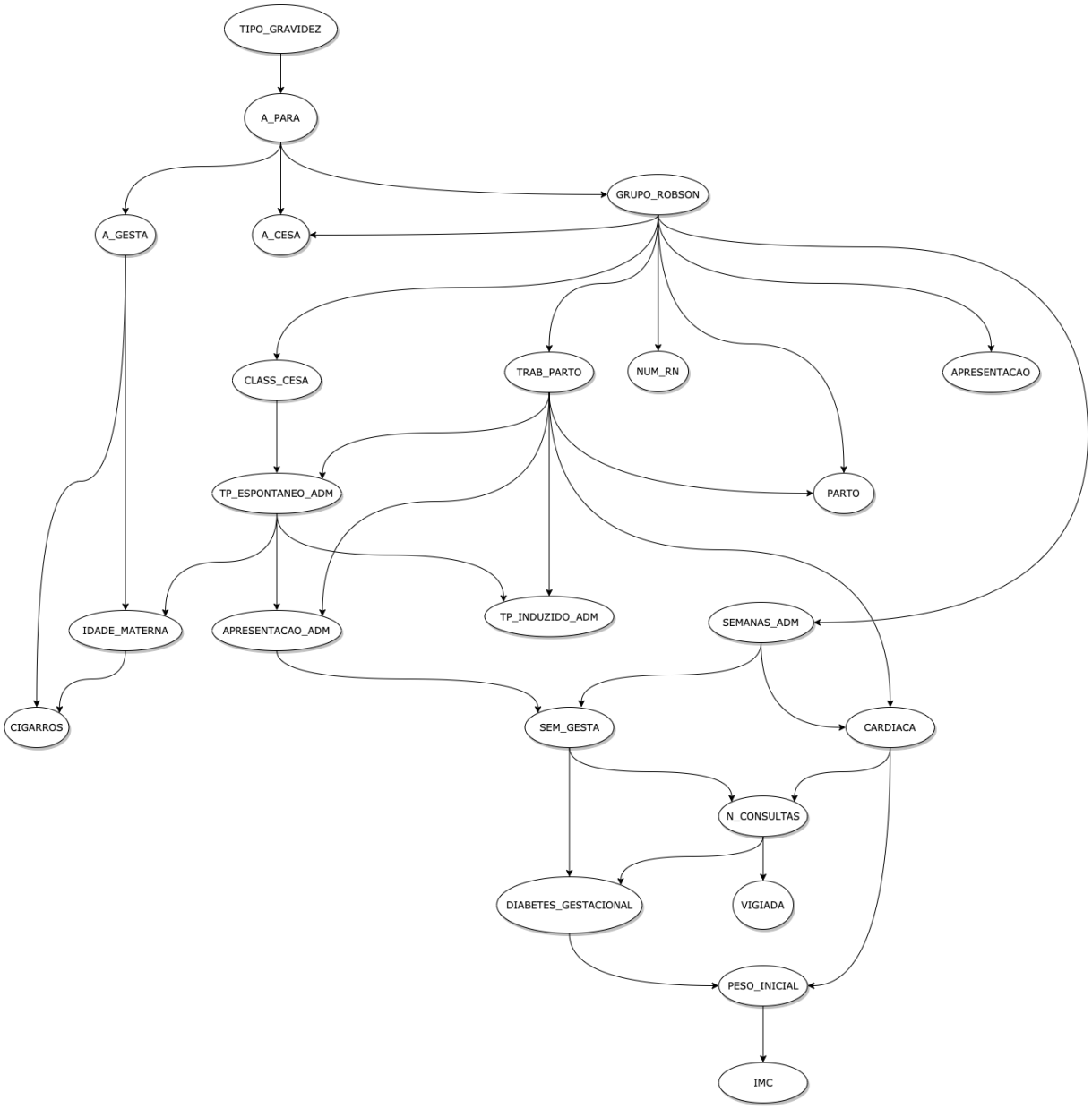


Figura 18 – Modelo gerado pelos dados do hospital H3 preenchidos com média e moda das variáveis



6.3 Detecção de anomalia

A comparação entre os 33 modelos criados pelos dados preenchidos com a distância de Hamming resultou em uma matriz 33x33 nula. As estruturas e relações geradas pelos dados preenchidos nos diferentes hospitais são idênticas.

Os modelos que apresentaram nodos excluídos da análise por não comportar dado algum na criação da Rede Bayesiana, como o modelo gerado pela seleção dos registros completos dos dados reais do H3, foram considerados uma anomalia pois sua estrutura e relação com os dados é tão diferente da maioria dos modelos criados que não podem ser comparados aos demais pela distância calculada. Todos outros modelos apresentaram mesma quantidade de nodos.

A estes outros modelos foram calculadas as distâncias de Hamming e construída a matriz de dissimilaridade. Foram incluídos 22 modelos criados a partir dos dados reais completos para encontrar as dissemelhanças de estruturas. Agregando as 22 entidades criou-se a matriz de dissimilaridade com 55 linhas e 55 colunas comparando entidade a entidade. Nesta comparação foram encontradas dissimilaridades diferentes de zero.

Na matriz de dissimilaridade todas comparações foram nulas, exceto com o modelo criado a partir da seleção dos registros completos do hospital H2, que apresentou $distnciaHamming = 16$ a todos outros. Ao executar os algoritmos de detecção de anomalia, e até mesmo em uma simples análise da matriz de dissimilaridade produzida, foi encontrado um modelo que destoa completamente dos outros. Não foi necessário classificar ou analisar o *score*, pois foi o único modelo destoante e por isso foi considerado anômalo. Visto que, 55 entidades comparadas, apenas uma entidade distanciou-se significativamente do padrão, considera-se que, do conjunto de simulação dos dados de hospitais, 1,81% teve comportamento anômalo.

Foi aplicado o questionário (Anexo A) apresentando as Tabelas 18 e 19 sobre a distribuição dos partos nos grupos de Robson nos dados que geraram os modelos considerados anômalos, apresentadas as taxas gerais de cesarianas e a distribuição dos tipos de parto pelos grupos de Robson para a especialista.

Segundo as respostas dadas pela especialista, na escala Likert, ela concordou em parte que ambas as entidades são anômalas. Para a anomalia referida na Tabela 18 foi dito que "a taxa de cesariana global está dentro do habitual. Algumas anomalias pontuais, por exemplo, não tem partos pélvicos em nulíparas (categoria 6) e os partos pélvicos são 100% cesariana nas multíparas o que é anômalo. Apresenta uma altíssima taxa de cesariana em partos gemelares (categoria 8).". Já para a anomalia referida na Tabela 19 foi dito "baixa taxa de parto por cesariana comparativamente com a média nacional, de resto todos os valores me parecem adequados para uma maternidade pequena.".

Tabela 18 – Distribuição da participação dos partos nos grupos de Robson dos dados que geraram o primeiro modelo considerado anômalo pela estrutura do modelo gerado conter diferentes nodos

GRUPO_ROBSON	**PARTO**			**Total**
	Cesariana	**Parto instrumentado**	**Parto eutócico**	
1	404 (20,1%)	715 (35,6%)	890 (44,3%)	2.009 (31%)
2	495 (38,2%)	408 (31,5%)	394 (30,4%)	1.297 (20%)
3	93 (5,9%)	172 (10,9%)	1.306 (83,1%)	1.571 (24%)
4	242 (30,9%)	80 (10,2%)	460 (58,8%)	782 (12%)
5	389 (50,6%)	177 (23,0%)	203 (26,4%)	769 (12%)
7	41 (100%)	0 (0%)	0 (0%)	41 (0.6%)
8	35 (89,7%)	4 (10,3%)	0 (0%)	39 (0.6%)
10	47 (73,4%)	7 (10,9%)	10 (15,6%)	64 (1.0%)
Total	1.746 (27%)	1.563 (24%)	3.263 (50%)	6.572 (100%)

Tabela 19 – Distribuição da participação dos partos nos grupos de Robson dos dados que geraram o modelo considerado anômalo pela detecção de anomalia

GRUPO_ROBSON	**PARTO**			**Total**
	Cesariana	**Parto instrumentado**	**Parto eutócico**	
1	142 (12,7%)	366 (32,9%)	606 (54,4%)	1.114 (28%)
2	251 (32,5%)	227 (29,4%)	294 (38,1%)	772 (20%)
3	26 (2,7%)	88 (9,0%)	867 (88,4%)	981 (25%)
4	39 (9,4%)	43 (10,4%)	331 (80,1%)	413 (11%)
5	301 (56,2%)	97 (18,1%)	138 (25,7%)	536 (14%)
6	12 (100%)	0 (0%)	0 (0%)	12 (0.3%)
7	31 (88,6%)	0 (0%)	4 (11,4%)	35 (0.9%)
8	11 (57,9%)	3 (15,8%)	5 (26,3%)	19 (0.5%)
9	4 (100%)	0 (0%)	0 (0%)	4 (0.1%)
10	20 (55,6%)	2 (5,6%)	14 (38,9%)	36 (0.9%)
Total	837 (21%)	826 (21%)	2.259 (58%)	3.922 (100%)

CAPÍTULO

7

CONCLUSÃO E CONTRIBUIÇÕES

"Nunca estou realmente satisfeita quanto a entender alguma coisa; porque, até onde entendo, a minha compreensão só pode ser uma fração infinitesimal de tudo o que eu quero compreender."
(Ada Lovelace, 1815-1852)

O trabalho apresentado corresponde a uma alternativa que se apresentou favorável para suprir a limitação do compartilhamento dos dados entre instituições e com a comunidade acadêmica. Apesar disto, o trabalho apossui limitações. Sobre o contexto do caso de estudo e momento histórico, o trabalho teve limitações de amostra e validação das anomalias encontradas. Apesar da intenção de mais hospitais e especialistas em integrarem o estudo, a amostra dos hospitais corresponde a uma parcela dos hospitais do norte de Portugal. Devido a pandemia vivida atualmente, a pesquisa de campo não foi realizada em sua totalidade e por isso a validação foi limitada por esta situação.

Apesar das limitações, foi possível contribuir com a construção do modelo representativo para cada hospital integrante no estudo. Adotando os métodos e *guidelines* realizados nesta pesquisa pode-se também comparar instituições de diversos contextos com problemas similares, o que é extremamente encorajado pela autora.

Para definir os modelos, compará-los e identificar as anomalias, primeiramente foram descritos detalhadamente todos os atributos contidos no conjunto de dados do caso de estudo. Junto aos especialistas nos dados e obstetra, foram encontradas divergências em uma análise criteriosa dos dados realizada. Foram validados e imputados dados conforme informações agregadas aos grupos de Robson identificados nos registros. Foram diminuídas as divergências encontradas nos dados e retirados os erros da coleta. Estas verificações e validações foram encaminhadas à empresa que realiza a coleta de dados.

A partir das verificações e validações dos valores contidos nos conjuntos de dados, foi realizado o preenchimento determinístico diminuindo inconsistência dos dados e aumentando a completude da informação nos conjuntos de dados. Também foram descritas as taxas de valores faltantes mesmo após este preenchimento. Com isto, concluímos que agregar o preenchimento automático em sistemas auxilia a análise, principalmente aos não especialistas do contexto dos dados. Apesar do preenchimento determinístico, notou-se que, em algumas variáveis, não há preenchimento ou não é realizado. É o caso das variáveis relativas à estimativa de peso nas ecografias semanais e posição do feto da 24^a a 42ª semana. Com os dados validados pelos especialistas foi realizada a análise da significância das variáveis para a definição dos tipos de parto. As variáveis com significância foram selecionadas para o estudo dos modelos.

Foram construídos modelos de rede Bayesiana para cada hospital. Para isso, os dados completos das variáveis relevantes para o tipo de participante foram selecionados. Viu se que, ao selecionar apenas dados completos, perde-se uma quantidade significativa de informação que muitas vezes gera impacto nas medidas estatísticas das variáveis e na distribuição dos registros em uma análise aplicada.

Por meio dos mecanismos que geraram os valores faltantes é possível identificar os tipos de preenchimento de dados que melhor representariam os dados reais. Os valores faltantes das variáveis dos conjuntos de dados dos hospitais foram preenchidos pelos de média e moda e pela técnica MICE. Os resultados na distribuição e medidas estatísticas de cada técnica de preenchimento foi semelhante e, por isso, conclui-se que o preenchimento com média/moda das variáveis seria suficiente para manter as características dos dados no contexto específico do caso de estudo. Foram gerados modelos dos dados imputados e ambos resultaram em estrutura igual. Sendo assim, as duas técnicas de preenchimento geraram resultados idênticos quanto a estrutura de rede Bayesiana gerada.

Com o novo conjunto de dados completos pelo preenchimento de dados, foi criado um modelo para cada hospital. Os modelos com base nos dados imputados serviram para gerar novos conjuntos de dados. Ao todo, 33 novos conjuntos foram criados. A matriz de dissimilaridade entre os 33 modelos resultou nula, constatando que a estrutura e as relações entre as características dos hospitais são iguais. Ao identificar tal situação, conclui-se que ao criar modelos com os dados recriados dos modelos dos dados originais, perdeu-se a singularidade de cada hospital. Nota-se isso pois, ao comparar com os modelos dos dados reais há nodos com diferentes relações e estas diferenças não foram transmitidas ao criar modelos pelos dados recriados. A integração de dados reais de mais hospitais e prevenção da comparação dos modelos pelos dados recriados para simulação de uma amostra de quantidade aproximada é relevante para o avanço da pesquisa e identificação de anomalias entre os modelos.

Para identificar novas comparações e simular outros hospitais participantes da compa-

ração, foi visto que o conjunto de dados na seleção dos registros completos não imputados dos hospitais apresentaram medidas estatísticas significativamente diferentes às dos dados reais e imputados. Foram incluídas na matriz de dissimilaridade a mesma proporção dos modelos (11 modelos por conjunto de dados, resultando na comparação de 55 modelos). Visto que um hospital não recriou todos os nodos como nos outros dois, este foi excluído desta nova análise e considerado como anomalia.

As dissimilaridades foram grandes de todas entidades quanto a uma entidade em específico. Todos os algoritmos definidos no capítulo de detecção de anomalias, especificamente LOF e KNN, identificaram a mesma observação como anômala, classificando todas as outras como não anômalas.

Sobre as duas entidades classificadas como anômalas, foram retirados dados escolhidos pela especialista do contexto médico para a sua análise. A distribuição dos partos pelos grupos de Robson, assim como as taxas dos tipos de partos, foram incluídas no questionário de validação. A totalidade das respostas sobre as duas entidades anômalas apresentadas à especialista foi "concordo em parte". Portanto, conclui-se que as anomalias identificadas pela técnica e análise de dados representam entidades anômalas pela comparação de modelos. No entanto, como os dados não dizem apenas sobre tipos de parto e grupos de Robson, a integridade dos dados não foi analisada pela especialista e esta poderia concluir por outras variáveis a definição das entidades como anômalas, concordando totalmente com a identificação.

Com o que foi apresentado, de acordo com as perguntas iniciadas que guiaram esta tese, acredita-se que as questões de pesquisa foram respondidas. Sendo assim, foi capacitada a comparação entre entidades por meio dos modelos com dados recriados com significância aos dados reais sem propriamente utilizá-los na comparação. E por meio desses dados a identificação de anomalias relevantes ao contexto do especialista foram identificadas e validadas sem a partilha dos dados entre as entidades.

Isso só foi possível pois os dados criados através dos modelos de redes Bayesianas representaram os dados originais, o que permitiu a comparação de diferentes fontes por meio dos modelos representativos e a identificação daqueles significativamente destoantes aos demais com relevância para a área específica dos dados. Por fim, o objetivo geral de detectar entidades anômalas pela comparação sem compartilhar dados entre elas foi alcançado.

7.1 Contribuições

A principal contribuição do trabalho é a possibilidade de realizar análises distribuídas de detecção de anomalia entre entidades "não comunicantes"sem a utilização dos dados reais. A detecção de anomalia por comparação de redes Bayesianas permitiu realizar a comparação entre

entidades sem a partilha de dados entre elas.

Ao conviver e participar da coleta de dados, soube da dificuldade dos hospitais em manterem uma motivação de preencher os dados. A rotina clínica não permite pausas para um preenchimento detalhado e muitas validações podem ser feitas em relação ao sistema que coleta os dados. As verificações e validações foram encaminhadas à empresa que realiza a coleta de dados. Além disso, foi proposto para que as imputações determinísticas tornem-se automáticas no preenchimento do sistema.

7.1.1 Publicações aceitas

Gelatti, G. J., de Carvalho, A. P., & Rodrigues, P. P. Detectar anomalias a fim de facilitar o diagnóstico médico por meio de agrupamento de dados e abstrações temporais. XXIV Jornadas de Classificação e Análise de Dados (JOCLAD 2017)., 97.

Gelatti, G. J., de Carvalho, A. C., & Rodrigues, P. P. (2017, June). Anomaly detection through temporal abstractions on intensive care data: position paper. In 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS) (pp. 354-355). IEEE.

Gelatti, G. J., Rodrigues, P. P., & Correia, R. J. C. (2020). Informatics as Support for Changes in Health Policy: A Case in Obstetrics. In HEALTHINF (pp. 745-749).

7.1.2 Publicação aguardando revisão

Gelatti, G. J., Maranhão, P., Reis, Z. & Correia, R. J. C. (2021). Evaluation of the impact of an electronic medical record in the use of Robson classification on cesarean section rates in Portuguese hospitals. International Journal of Medical Informatics.

7.2 Trabalhos futuros

A partir do caso de estudo identifica-se a necessidade de investigar as relações dos modelos gerados com a previsão do tipo de parto e descrição da classificação das gestantes. Acredita-se que as novas relações identificadas pelos modelos produzidos com características não incluídas na Classificação de Robson (CR) podem agregar valor à classificação.

Algumas questões foram levantadas no desenvolvimento deste estudo:

Qual a influência de diferentes medidas de distância ou (dis)similaridade ao comparar modelos levando em consideração as probabilidades dos nodos?

É possível comparar modelos com quantidade de nodos diferentes?

Existe variação significativa no resultado quanto a escolha da medidas de dissimilaridade entre grafos ou modelos e as técnicas de detecção de anomalias? O resultado dessas escolhas variam conforme o contexto da aplicação?

Acredita-se que estas são perguntas de grande valor para a continuação de novas pesquisas neste contexto. Haverá sempre descobertas e melhorias a serem feitas nas técnicas e aplicações do trabalho. Destaco o envolvimento direto desta tese com o contexto clínico e hospitalar, principalmente na troca de conhecimentos entre áreas e retorno direto da pesquisa à sociedade.

REFERÊNCIAS

151/2019, S. I. d. .-.-. Diário da República n.º. Lei n.º 58/2019. Disponível em: <https://data.dre.pt/eli/lei/58/2019/08/08/p/dre>. Citado na página 28.

ACID, S.; CAMPOS, L. M. de. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. **Journal of Artificial Intelligence Research**, v. 18, p. 445–490, 2003. Citado na página 48.

AGGARWAL, C. C. **Outlier Analysis**. Springer, 2013. ISBN 978-1-4614-6395-5. Disponível em: <http://dx.doi.org/10.1007/978-1-4614-6396-2>. Citado nas páginas 43 e 45.

ANTAL, P.; FANNES, G.; TIMMERMAN, D.; MOREAU, Y.; MOOR, B. D. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. **Artificial Intelligence in Medicine**, v. 29, n. 1, p. 39 – 60, 2003. ISSN "0933-3657". Artificial Intelligence in Medicine Europe AIME '01. Citado na página 33.

AYER, T.; CHHATWAL, J.; ALAGOZ, O.; KAHN, J. C. E.; WOODS, R. W.; BURNSIDE, E. S. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. **RadioGraphics**, v. 30, n. 1, p. 13–22, 2010. PMID: 19901087. Citado na página 33.

BREUNIG, M. M.; KRIEGEL, H.-P.; NG, R. T.; SANDER, J. Lof: identifying density-based local outliers. In: **Proceedings of the 2000 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 2000. p. 93–104. Citado na página 48.

BRUNO, G.; GARZA, P. Tod: Temporal outlier detection by using quasi-functional temporal dependencies. **Data & Knowledge Engineering**, Elsevier, v. 69, n. 6, p. 619–639, 2010. Citado na página 45.

BUUREN, S. V. **Flexible imputation of missing data**. [S.l.]: CRC press, 2018. Citado na página 53.

BUUREN, S. V.; OUDSHOORN, K. **Flexible multivariate imputation by MICE**. [S.l.]: Leiden: TNO, 1999. Citado na página 54.

BUUREN, S. van; GROOTHUIS-OUDSHOORN, K.; ROBITZSCH, A.; VINK, G.; DOOVE, L.; JOLANI, S. Package 'mice'. **Computer software**, 2015. Citado na página 54.

CAMPOS, G. O.; ZIMEK, A.; SANDER, J.; CAMPELLO, R. J.; MICENKOVÁ, B.; SCHUBERT, E.; ASSENT, I.; HOULE, M. E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. **Data mining and knowledge discovery**, Springer, v. 30, n. 4, p. 891–927, 2016. Citado na página 49.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, ACM, v. 41, n. 3, p. 15, 2009. Citado nas páginas 15, 25, 43, 44, 45 e 49.

CHRISTENSEN, R. Testing fisher, neyman, pearson, and bayes. **The American Statistician**, Taylor & Francis, v. 59, n. 2, p. 121–126, 2005. Citado na página 73.

CRUZ-CORREIA, R.; AMORIM, E. Obs.care: aplicacao de cuidados intensivos de obstetricia. **Porto: Universidade do Porto/Faculdade de Medicina/Departamento Ciencias da Informacao e da Decisao em Saude**, 2008. Citado na página 61.

DARWICHE, A. Bayesian networks. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 12, p. 80–90, dez. 2010. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/1859204.1859227>. Citado nas páginas 40 e 55.

DIAS. **DiAS: De-identification as a Service**. 2018. Disponível em: <http://hltsys.pt/pt/projeto_dias/>. Citado na página 52.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. Citado nas páginas 47 e 48.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. Inteligência artificial–uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, 2011. Citado na página 43.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine learning**, Springer, v. 29, n. 2-3, p. 131–163, 1997. Citado na página 29.

GELATTI, G. J.; RODRIGUES, P. P.; CORREIA, R. J. C. Informatics as support for changes in health policy: A case in obstetrics. In: **HEALTHINF**. [S.l.: s.n.], 2020. p. 745–749. Citado na página 68.

GIERE, R. N. How models are used to represent reality. **Philosophy of Science**, v. 71, n. 5, p. 742–752, 2004. Disponível em: <https://doi.org/10.1086/425063>. Citado na página 33.

HADI, A. S. Identifying multiple outliers in multivariate data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 54, n. 3, p. 761–771, 1992. Citado na página 45.

HARPER, L. M.; ODIBO, A. O.; MACONES, G. A.; CAHILL, A. G. Predicting cesarean in the second stage of labor. **American journal of perinatology**, Thieme Medical Publishers, v. 30, n. 10, p. 827–832, 2013. Citado na página 33.

HAWKINS, D. M. **Identification of outliers**. [S.l.]: Springer, 1980. v. 11. Citado nas páginas 25 e 43.

HOUSSEINI, A. A.; NEWMAN, T.; COX, A.; DEVOE, L. D. Prediction of risk for cesarean delivery in term nulliparas: a comparison of neural network and multiple logistic regression models. **American Journal of Obstetrics & Gynecology**, Elsevier, v. 201, n. 1, p. 113–e1, 2009. Citado na página 33.

HUTSON, M. **Artificial intelligence faces reproducibility crisis**. [S.l.]: American Association for the Advancement of Science, 2018. Citado na página 29.

KOTASKA, A. Caesarean section or vaginal delivery in the 21st century. **Entre Nous: The Eropean magazine for sexual and reproductive health**, Division of Noncommunicable Diseases and Promoting Health through the Life-course, n. 81, p. 8–9, 2015. Citado na página 59.

LAURITZEN, S. L. **Graphical models**. [S.l.]: Clarendon Press, 1996. v. 17. Citado na página 33.

LI, Y. Anomaly detection in unknown environments using wireless sensor networks. 2010. Citado na página 43.

LIKERT, R. A technique for the measurement of attitudes. **Archives of psychology**, 1932. Citado nas páginas 52 e 57.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2019. v. 793. Citado na página 54.

LOVELL, D.; ROSARIO, B.; NIRANJAN, M.; PRAGER, R.; DALTON, K.; DEROM, R.; CHALMERS, J. Design, construction and evaluation of systems to predict risk in obstetrics. **International journal of medical informatics**, Elsevier, v. 46, n. 3, p. 159–173, 1997. Citado na página 33.

MARDY, A. H.; ANANTH, C. V.; GROBMAN, W. A.; GYAMFI-BANNERMAN, C. A prediction model of vaginal birth after cesarean in the preterm period. **American Journal of Obstetrics & Gynecology**, Elsevier, v. 215, n. 4, p. 513–e1, 2016. Citado na página 33.

METZ, T. D.; STODDARD, G. J.; HENRY, E.; JACKSON, M.; HOLMGREN, C.; ESPLIN, S. Simple, validated vaginal birth after cesarean delivery prediction model for use at the time of admission. **Obstetrics and gynecology**, NIH Public Access, v. 122, n. 3, p. 571, 2013. Citado na página 33.

MURRAY, R. S. Theory and problems of probability and statistics. **Singapore: McGraws-Hill Book and Co**, 1975. Citado na página 35.

MYLONAS, I.; FRIESE, K. Indications for and risks of elective cesarean section. **Deutsches Ärzteblatt International**, Deutscher Arzte-Verlag GmbH, v. 112, n. 29-30, p. 489, 2015. Citado nas páginas 26 e 59.

NEAL, R. M. Priors for infinite networks. In: ____. **Bayesian Learning for Neural Networks**. New York, NY: Springer New York, 1996. p. 29–53. ISBN 978-1-4612-0745-0. Disponível em: <https://doi.org/10.1007/978-1-4612-0745-0_2>. Citado na página 40.

NEAPOLITAN, R. E. *et al.* **Learning bayesian networks**. [S.l.]: Pearson Prentice Hall Upper Saddle River, NJ, 2004. v. 38. Citado nas páginas 34 e 54.

ORGANIZATION, W. H. *et al.* **WHO statement on caesarean section rates**. [S.l.], 2015. Citado na página 59.

PEARL, J. **Probabilistic inference in intelligent systems**. [S.l.]: Morgan Kaufmann San Mateo, CA, 1988. Citado na página 33.

PEREIRA, R. C.; SANTOS, M. S.; RODRIGUES, P. P.; ABREU, P. H. Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. **Journal of Artificial Intelligence Research**, v. 69, p. 1255–1285, 2020. Citado na página 54.

PERRIER, E.; IMOTO, S.; MIYANO, S. Finding optimal bayesian network given a super-structure. **Journal of Machine Learning Research**, v. 9, n. Oct, p. 2251–2286, 2008. Citado na página 48.

PISANI, E.; ABOUZAHR, C. Sharing health data: good intentions are not enough. **Bulletin of the World Health Organization**, SciELO Public Health, v. 88, n. 6, p. 462–466, 2010. Citado na página 29.

PORTUGAL, E. **Instituto Nacional de Estatística - Estatísticas da Saúde : 2018.** [s.n.], 2020. ISSN 2183-1637. ISBN 978-989-25-0534-3. Disponível em: <https://www.ine.pt/xurl/pub/257793024>>. Citado na página 82.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient algorithms for mining outliers from large data sets. In: **Proceedings of the 2000 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 2000. p. 427–438. Citado na página 47.

RAY, A.; JOSE, S. Analysis of caesarean-section rates according to robson's ten group classification system and evaluating the indications within the groups. **Int J Reprod Contracept Obstet Gynecol**, v. 6, n. 2, p. 447–51, 2017. Citado na página 67.

RAYANA, S.; AKOGLU, L. Less is more: Building selective anomaly ensembles. **Acm transactions on knowledge discovery from data (tkdd)**, ACM New York, NY, USA, v. 10, n. 4, p. 1–33, 2016. Citado na página 43.

REGULAMENTO Geral sobre a Proteção de Dados. Regulamento (UE) 2016/679 do parlamento europeu e do conselho de 27 de abril de 2016 relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE, 2016. Disponível em: <https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=EN>. Citado na página 28.

ROBSON, M. S. Classification of caesarean sections. **Fetal and maternal medicine review**, v. 12, n. 1, p. 23–39, 2001. Citado nas páginas 60, 63 e 76.

ROYSTON, P.; WHITE, I. R. *et al.* Multiple imputation by chained equations (mice): implementation in stata. **J Stat Softw**, v. 45, n. 4, p. 1–20, 2011. Citado na página 83.

RUBIN, D. B. Inference and missing data. **Biometrika**, v. 63, n. 3, p. 581–592, 12 1976. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/63.3.581>. Citado na página 53.

SCHUBERT, E.; ZIMEK, A.; KRIEGEL, H.-P. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. **Data mining and knowledge discovery**, Springer, v. 28, n. 1, p. 190–237, 2014. Citado na página 49.

SCUTARI, M.; SCUTARI, M. M.; MMPC, H.-P. **Package 'bnlearn'**. 2020. Citado na página 55.

SILVER, R. M.; LANDON, M. B.; ROUSE, D. J.; LEVENO, K. J.; SPONG, C. Y.; THOM, E. A.; MOAWAD, A. H.; CARITIS, S. N.; HARPER, M.; WAPNER, R. J. *et al.* Maternal morbidity associated with multiple repeat cesarean deliveries. **Obstetrics & Gynecology**, LWW, v. 107, n. 6, p. 1226–1232, 2006. Citado na página 26.

SMITH, G. C.; DELLENS, M.; WHITE, I. R.; PELL, J. P. Combined logistic and bayesian modeling of cesarean section risk. **American Journal of Obstetrics & Gynecology**, Elsevier, v. 191, n. 6, p. 2029–2034, 2004. Citado na página 33.

SUN, F.; DENG, Y.; DENG, F.; ZHU, Q.; CHU, H. Unsupervised maritime traffic pattern extraction from spatio-temporal data. In: IEEE. **2015 11th International Conference on Natural Computation (ICNC)**. [S.l.], 2015. p. 1218–1223. Citado na página 45.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *et al.* **Introduction to data mining**. [S.l.]: Pearson Addison Wesley Boston, 2006. v. 1. Citado nas páginas 27, 43, 44 e 45.

TSAMARDINOS, I.; BROWN, L. E.; ALIFERIS, C. F. The max-min hill-climbing bayesian network structure learning algorithm. **Machine learning**, Springer, v. 65, n. 1, p. 31–78, 2006. Citado na página 48.

UNIÃO SEÇÃO 1, p. . Diário Oficial da. Lei nº 13.709, de 14 de agosto de 2018. 2018. Citado na página 29.

WALPORT, M.; BREST, P. Sharing research data to improve public health. **The Lancet**, Elsevier, v. 377, n. 9765, p. 537–539, 2011. Citado na página 29.

WICHERTS, J. M.; BAKKER, M.; MOLENAAR, D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. **PloS one**, Public Library of Science, v. 6, n. 11, p. e26828, 2011. Citado na página 29.

WU, S. X.; BANZHAF, W. The use of computational intelligence in intrusion detection systems: A review. **Applied Soft Computing**, Elsevier, v. 10, n. 1, p. 1–35, 2010. Citado na página 27.

YIN, S.-N.; KANG, H.-S.; KIM, S.-R. Clustering algorithm based on outlier detection for anomaly intrusion detection. **Internet Technology Journal**, Airiti Press Inc., & Ainosco Press, v. 17, n. 2, p. 291–299, 2016. Citado na página 43.

GLOSSÁRIO

Tabu search (TABU): Algoritmo de busca local. A técnica consiste na exploração contínua, dentro de um espaço de busca, a partir de uma solução inicial tentando avançar para uma outra solução na sua vizinhança até que se satisfaça um determinado critério de parada.

***Cross table*:** Tabela de contingência que compara a distribuição de valores em duas variáveis.

***Hill-climbing* (HC):** Técnica de otimização matemática. Algoritmo guloso: inicia uma solução arbitrária para um problema e tenta encontrar uma solução melhor fazendo alterações incrementais.

Classificação de Robson: A classificação de Robson é amplamente utilizada para análise das taxas de cesariana em Portugal. Ela classifica as grávidas em 10 grupos.

Especialista nos dados: Pessoa que manipula ou coleta os dados.

ESPONTANEA: Variável relativa a gravidez espontânea. Gravidez sem tratamentos associados.

ESTIMULACAO: Variável relativa a gravidez com utilização de métodos para a estimulação.

FIV: Variável relativa a gravidez iniciada após estimulação por tratamento de fertilização in vitro.

ICSI: Variável relativa a gravidez iniciada após estimulação por tratamento com aplicação de injeção intra-citoplásmica.

INSEMINACAO: Variável relativa a gravidez iniciada após estimulação por inseminação artificial.

Parto auxiliado: Também chamado de parto instrumentado, é o tipo de parto realizado com auxílio de instrumentos. Os instrumentos mais utilizadas para auxiliar o parto são ventosa, fórceps ou espátula.

RAI: O responsável pelo acesso a informação.

TEC: Variável relativa a gravidez iniciada após estimulação por transferência de embriões crio-preservados.

APÊNDICE

A

QUESTIONÁRIO DE VALIDAÇÃO

Questionário aplicado a obstetra sobre as anomalias encontradas.

Avaliação das anomalias identificadas por meio da comparação de modelos

O objetivo deste questionário é a validação do estudo feito por Giovana Jaskulski Gelatti resultado da tese de doutorado intitulada "Detecção de anomalia através da comparação de modelos representativos" do Programa Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC) da Universidade de São Paulo – Brasil com orientação do Professor Dr. André Ponce de Leon Ferreira de Carvalho. O estudo é um coorte de informação sobre 26.020 grávidas de 3 secções de obstetrícia de hospitais do norte de Portugal entre os anos de 2015 a 2020. As variáveis utilizadas foram sobre o contexto obstétrico referente a gestação e histórico gestacional, assim como informações sobre a supervisão da gestação e histórico de doenças.

Instruções: É considerado uma anomalia o comportamento que difere substancialmente dos outros. Neste contexto, pretende-se avaliar as anomalias identificadas por meio das informações referentes às taxas gerais nos tipos de parto e distribuição dos partos nos grupos da classificação de Robson descritos no Quadro 1. Para cada tabela, responder se, a partir dos dados apresentados, considera a secção de obstetrícia diferente do padrão de sua experiência e conhecimento. Ao responder o questionário, declara-se concordante com a utilização e publicação dos dados aqui recolhidos.

Quadro 1 – Descrição dos grupos da classificação de Robson

Grupo	Descrição
1	Mulheres nulíparas, único feto em posição cefálica, com semanas de gestação maior ou igual a 37 semanas em trabalho de parto espontâneo
2	Mulheres nulíparas, único feto, em posição cefálica, maior ou igual a 37 semanas de gestação que tiveram trabalho de parto induzido ou cesariana antes do trabalho de parto
3	Mulheres multíparas, sem cesariana anterior, único feto, em posição cefálica maior ou igual a 37 semanas em trabalho de parto espontâneo
4	Mulheres multíparas, sem cesariana anterior, único feto, em posição cefálica maior ou igual a 37 semanas que tiveram trabalho de parto induzido ou cesárea
5	Todas as mulheres multíparas, com pelo menos uma cesariana anterior , único feto , em posição cefálica com maior ou igual a 37 semanas de gestação
6	Todas as mulheres nulíparas, único feto, posição pélvica
7	Todas as mulheres multíparas, único feto, posição pélvica
8	Todas as mulheres com gestações múltiplas
9	Todas as mulheres com único feto em posição transversal ou oblíqua
10	Todas as mulheres com único feto, posição cefálica com menos ou igual a 36 semanas de gestação

Questionário

Tabela 1 - Dados do Hospital 1 referentes à classificação de Robson e participação nos tipos de parto no período 10/2016 - 11/2020.

	PARTO			
GRUPO_ROBSON	**Cesariana**	**Parto instrumentado**	**Parto eutócico**	**Total**
1	404 (20,1%)	715 (35,6%)	890 (44,3%)	2,009 (31%)
2	495 (38,2%)	408 (31,5%)	394 (30,4%)	1,297 (20%)
3	93 (5,9%)	172 (10,9%)	1,306 (83,1%)	1,571 (24%)
4	242 (30,9%)	80 (10,2%)	460 (58,8%)	782 (12%)
5	389 (50,6%)	177 (23,0%)	203 (26,4%)	769 (12%)
7	41 (100%)	0 (0%)	0 (0%)	41 (0.6%)
8	35 (89,7%)	4 (10,3%)	0 (0%)	39 (0.6%)
10	47 (73,4%)	7 (10,9%)	10 (15,6%)	64 (1.0%)
Total	1,746 (27%)	1,563 (24%)	3,263 (50%)	6,572 (100%)

1. De acordo com os dados apresentados na Tabela 1, você considera este hospital uma anomalia?

discordo plenamente	discordo em parte	nem concordo nem discordo	concordo em parte	concordo plenamente
☐	☐	☐	x	☐

Espaço para argumentação (não obrigatório):

A taxa de cesariana global está dentro do habitual.
Algumas anomalias pontuais, por exemplo, não tem partos pélvicos em nulíparas (categoria 6) e os partos pélvicos são 100% cesariana nas multíparas o que é anómalo. Apresenta uma altíssima taxa de cesariana em partos gemelares (categoria 8).

Tabela 2 - Dados do Hospital 2 referentes à classificação de Robson e participação nos tipos de parto no período 03/2015 - 11/2020.

	PARTO			
GRUPO_ROBSON	**Cesariana**	**Parto instrumentado**	**Parto eutócico**	**Total**
1	142 (12,7%)	366 (32,9%)	606 (54,4%)	1,114 (28%)
2	251 (32,5%)	227 (29,4%)	294 (38,1%)	772 (20%)
3	26 (2,7%)	88 (9,0%)	867 (88,4%)	981 (25%)
4	39 (9,4%)	43 (10,4%)	331 (80,1%)	413 (11%)
5	301 (56,2%)	97 (18,1%)	138 (25,7%)	536 (14%)
6	12 (100%)	0 (0%)	0 (0%)	12 (0.3%)
7	31 (88,6%)	0 (0%)	4 (11,4%)	35 (0.9%)
8	11 (57,9%)	3 (15,8%)	5 (26,3%)	19 (0.5%)
9	4 (100%)	0 (0%)	0 (0%)	4 (0.1%)
10	20 (55,6%)	2 (5,6%)	14 (38,9%)	36 (0.9%)
Total	837 (21%)	826 (21%)	2,259 (58%)	3,922 (100%)

2. De acordo com os dados apresentados na Tabela 2, você considera este hospital uma anomalia?

discordo plenamente	discordo em parte	nem concordo nem discordo	concordo em parte	concordo plenamente
☐	☐	☐	x	☐

Espaço para argumentação (não obrigatório):

Baixa taxa de parto por cesariana comparativamente com a média nacional, de resto todos os valores me parecem adequados para uma maternidade pequena.

ANEXO

A

AUTORIZAÇÕES DOS COMITÊS DE ÉTICA

Autorizações dos comitês de ética dos hospitais HSO, CHEDV e CHVNGE. Parecer e autorização para a utilização dos dados requeridos ao comitê de ética do Hospital de Senhora da Oliveira (HSO).

REPÚBLICA PORTUGUESA SAÚDE | 40 ANOS | SNS SERVIÇO NACIONAL DE SAÚDE | Hospital da Senhora da Oliveira GUIMARÃES | CENTRO ACADÉMICO E DE FORMAÇÃO Hospital da Senhora da Oliveira Guimarães

PARECER DO COORDENADOR DO CENTRO ACADÉMICO

Título: *"Predição e análise do tipo de parto em gestantes portuguesas através de Redes Bayesianas."*

Ref: 85/2020 – Projeto de Investigação

Investigador Principal: Prof. Doutor Ricardo João Cruz Correia da Universidade do Porto.

Investigador Responsável no HSOG: Dr. Ricardo Santos, Serviço de Ginecologia e Obstetrícia do HSOG.

Avaliação da exequibilidade e de mérito científico: Estudo com interesse clínico e académico, pelo que, nada a opor ao presente projeto.

Com os melhores cumprimentos,

Pedro Guimarães Cunha

Prof. Doutor Pedro Guimarães Cunha
Coordenador do Centro Académico e de Formação do HSOG

CENTRO ACADÉMICO E DE FORMAÇÃO Hospital da Senhora da Oliveira Guimarães

PARECER DO COMISSÃO DE ÉTICA PARA A SAÚDE

Nos termos da reunião desta Comissão de Ética, dá-se o conhecimento a V. Exa. do parecer emitido em reunião extraordinária do dia 13 de Novembro de 2020:
Analisado o projeto de investigação, *"Predição e análise do tipo de parto em gestantes portuguesas através de Redes Bayesianas"*, que tem como Investigador Principal o Prof. Doutor Ricardo João Cruz Correia da Universidade do Porto e como Investigador Responsável no HSOG o Dr. Ricardo Santos do Serviço de Ginecologia e Obstetrícia, a Comissão de Ética deixa as seguintes considerações:

Nada a opor, desde que os dados clínicos sejam obtidos pelo investigador do HSO

Com os melhores cumprimentos,

Dr. João Lima Reis
Presidente da CES

Termos de Autorização Local

Utilize os modelos que forem aplicáveis ao Estudo em causa, acrescentando a informação que considerar pertinente.

Estudo de investigação (título)

Directores de Serviço

Na qualidade de Director de Serviço, declaro que autorizo a execução do Estudo de Investigação acima mencionado e comprometo-me a prestar as condições necessárias para a boa execução do mesmo, de acordo com o programa de trabalhos e os meios apresentados.

Serviço	Nome do Director	Data	Assinatura
Serviço de obstetrícia.	José Furtado.	16/10/2020	

Directores / Conselhos de Gestão de Departamento

Na qualidade de Director do Departamento, declaro que autorizo a execução do Estudo de Investigação acima mencionado e comprometo-me a prestar as condições necessárias para a boa execução do mesmo, de acordo com o programa de trabalhos e os meios apresentados.

Departamento	Nome do Director	Data	Assinatura
Serviço de obstetrícia	José Furtado.	16/10/2020	

Responsáveis por Unidades, Gabinetes ou Sectores (se aplicável)

Na qualidade de Responsável pela Unidade / Gabinete / Sector, dou parecer favorável à execução do Estudo de Investigação acima mencionado e comprometo-me a prestar as condições necessárias para a boa execução do mesmo, de acordo com o programa de trabalhos e os meios apresentados.

Unidade / Gabinete / Sector	Nome do Director	Data	Assinatura
_____	_____	__/__/____	_____
_____	_____	__/__/____	_____

Chefias Técnicas (se aplicável)

Na qualidade de Enfermeiro Chefe / Técnico Coordenador do Serviço abaixo indicado, dou parecer favorável à execução do Estudo de Investigação acima mencionado e comprometo-me a prestar as condições necessárias para a boa execução do mesmo, de acordo com o programa de trabalhos e os meios apresentados.

Serviço	Chefia Técnica	Data	Assinatura
_____	_____	__/__/____	_____
_____	_____	__/__/____	_____

Parecer e autorização para a utilização dos dados requeridos ao comitê de ética do Centro Hospitalar de entre o Douro e Vouga (CHEDV).

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

Centro Hospitalar de Entre o Douro e Vouga, E.P.E.

Exmo. Senhor
Dr. Ricardo João Cruz Correia

rcorreia@med.up.pt

SUA REFERÊNCIA	SUA COMUNICAÇÃO DE	NOSSA REFERÊNCIA	DATA:
		Nº: CA-371/2020-0t_MP/CC	22 de dezembro de 2020

ASSUNTO: "Predição e análise do tipo de parto em gestantes portuguesas através de redes bayesianas"

No seguimento do estudo identificado em epígrafe, submetido por V. Exa., informamos que a mesmo foi objeto de apreciação em reunião do Conselho de Administração de 22 de dezembro de 2020, tendo sido proferido o seguinte despacho:

"Deliberado autorizar."

Com os melhores cumprimentos,

O Presidente do Conselho de Administração

Miguel Paiva, Dr.

Centro Hospitalar de Entre o Douro e Vouga, E.P.E.
Rua Dr. Cândido Pinho, 4520-211 Santa Maria da Feira, PORTUGAL
TEL + 351 256 379 700 **EMAIL** geral@chedv.min-saude.pt **www.chedv.min-saude.pt**

Centro Hospitalar
de Entre o Douro e Vouga, E.P.E.

Á Comissão de Ética,

Declaro para os devidos fins que eu, Teresa Paula Teles, diretora do Serviço de obstetrícia do Centro Hospitalar de Entre Douro e Vouga (CHEDV), estou de acordo com a realização do estudo intitulado "Predição e análise do tipo de parto em gestantes portuguesas através de Redes Bayesianas", o qual apresenta como investigador principal o Prof. Dr. Ricardo J. Cruz Correia.

Sem mais no momento,

Com os meus melhores cumprimentos,

Porto, 13 de outubro de 2020.

A diretora de serviço,

Assinatura

Parecer e autorização para a utilização dos dados requeridos ao comitê de ética do Centro Hospitalar de Vila Nova de Gaia e Espinho (CHVNGE).

192

Foi submetido o formulário com a referencia: UIEC-2020-10159482865fa924b33b101

I.MU3

CHVNG/E 53038 11-11-'20

Titulo

Titulo do Projecto

Predição e análise do tipo de parto em gestantes portuguesas através de Redes Bayesianas

Investigador Principal

Centro Hospitalar Vila Nova de Gaia/Espinho, E.P.E

Nome completo

Ricardo J. Cruz Correia

Contacto telefónico

966316853

Endereço de Email

rcorreia@med.up.pt

Equipa de Investigação

Nome dos elementos da equipa

Pedro Pereira Rodrigues; Eliana Sousa; Giovana Gelatti; Mafalda Laranjo; Ana Paula Machado.

Afiliação dos elementos da Equipa

Faculdade de Medicina da Universidade do Porto, Unidade Local de Saúde de Matosinhos e Centro Hospitalar universitário São João

Nome do orientador

Autorizado
Diana Mota
2020/11/25

Diana Mota
Diretora Clínica

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

CENTRO HOSPITALAR VILA NOVA DE GAIA/ESPINHO
Comissão de Ética

PARECER DA COMISSÃO DE ÉTICA PARA A SAÚDE

TÍTULO DO ESTUDO/PEDIDO

"Predição e análise do tipo de parto em gestantes portuguesas através de Redes Bayesianas "

Documento do CES: 192/2020

Serviço onde irá decorrer o Estudo: Obstetrícia

Investigador Principal: Ricardo J. Cruz Correia

Centro Hospitalar Vila Nova de Gaia/Espinho, E.P.E

A Comissão de Ética para a Saúde do Centro Hospitalar de Vila Nova de Gaia/Espinho, EPE, em reunião ordinária do dia **19/11/2020**, apreciou a documentação constante do dossier submetido para o estudo acima referenciado:

- Formulário de submissão a UIEC;
- Formulário da CES;
- Pedido de realização do estudo de investigação dirigido ao CA;
- Aprovação do projeto de investigação pelo diretor de serviço;
- Compromisso de comunicação dos resultados obtidos com o projeto de investigação;
- Declaração de conflito de interesses
- Curriculum vitae do investigador principal;
- Curriculum vitae dos restantes elementos da equipe de investigação;
- Projeto do estudo.

Apreciação:

Estudo retrospetivo transversal de uma coorte de dados multicêntrico, a partir de dados anonimizados registados no registo eletrónico para obstetrícia denominado ObsCare .

Será realizado em 8 hospitais portugueses, que tiverem interesse na participação do estudo. Serão incluídas todas as gestantes atendidas desde o período de implementação do registo ate outubro de 2020 .

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

CENTRO HOSPITALAR VILA NOVA DE GAIA/ESPINHO

A analise dos dados será realizada desde o momento da implementação do ObsCare nesta instituição ate ao mês de outubro de 2020 .

A base de dados será protegida pela ferramenta DiAS – De-identification as a Service,ferramenta da empresa *HealthySystems* (HLTSYS).esta base num primeiro momento sofrerá a anonimização ,e em segundo momento será realizada a pseudo-anonimização através do DiAS .

Serão utilizadas a volta de 60 variáveis: demográficas e clinicas (dados não sensíveis).

O objetivo da investigação: desenvolver a partir do modelo estatístico das redes bayesianas (RBs), um sistema de apoio a decisão a fim de auxiliar, planejar e predizer o tipo de parto para as gestantes portuguesas .

O tempo previsto para a duração da investigação é de 13 meses.

Não está prevista a obtenção do consentimento informado.

Garantem confidencialidade e anonimato, na base de dados criada não constará identificação ou outro tipo de dados que possam permitir identificação.

A base de dados será encriptada e terá uma password de acesso somente do investigador principal. Para a utilização da base dados será necessário fazer copias os portáteis dos investigadores autorizados para analise. Os investigadores com autorização para utilização dos dados terão os portáteis com credencias diferenciadas, e os ficheiros encriptados com passwords de acesso diferentes por colaborador.

Os dados do estudo serão conservados ate 31/12/2021.

Não está prevista a transferência de dados.

Esta prevista a publicação dos resultados da investigação.

Não apresenta custos acrescidos para o participante ou instituição.

O trabalho não e é financiado.

Consideram não haver conflito de interesses.

Ouvido o relator, o processo foi votado pelos membros da Comissão de Ética para a Saúde do Centro Hospitalar de Vila Nova de Gaia/EPE presentes:

Presidente: Enfª Ana Saraiva

Vice-presidente: Drª Paula Fernandes

Restantes Membros:

Drª Ana Ferreira

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

CENTRO HOSPITALAR VILA NOVA DE GAIA/ESPINHO
Comissão de Ética

Drª Ana Isabel Paixão
Enfª Teresa Trigo

Delibera-se parecer favorável, o qual foi aprovado por unanimidade dos presentes.

Data: 19/11/ 2020

A Presidente da Comissão de Ética para a Saúde

Enf. Ana Saraiva

Centro Hospitalar Vila Nova de Gaia/Espinho, E.P.E

ICMC USP
SÃO CARLOS